# Evaluating Validity Properties of 25 Race-Related Scales

Neil Hester[1], Jordan R. Axt[1], Nellie Siemers[1], Eric Hehman[1]

[1]Department of Psychology, McGill University, Montreal, QC, H3A 1G1, Canada

***NOTE:*** *This is an unpublished manuscript in press at Behavior Research Methods, shared as a preprint to disseminate the work in a timely manner and encourage feedback. This is NOT a copy of record.*

**Corresponding Author:**

Neil Hester, Department of Psychology, McGill University, E-mail: neilrhester@gmail.com

**Author Contributions:**

Conceived research: N.H. Methodology: N.H., J.R.A., E.H. Data Curation: N.H., J.R.A., N.S. Data Collection: N.H., E.H. Analysis: N.H. Writing – Original Draft: N.H. Writing – Review and Editing: All authors.

**Repository:**

All data, scripts, figures, analysis markdowns, and other supplementary materials are available at https://osf.io/zg6fr/?view_only=e6d56172a5934259a81729312ebf0754.

Abstract

Racial attitudes, beliefs, and motivations lie at the center of many of the most influential theories of prejudice and discrimination. The extent to which such theories can meaningfully explain behavior hinges on accurate measurement of these latent constructs. We evaluated the validity properties of 25 race-related scales in a sample of 1,031,207 respondents using modern approaches such as dynamic fit indices, Item Response Theory, and nomological nets. Despite showing adequate internal reliability, many scales demonstrated poor model fit and had latent score distributions showing clear floor or ceiling effects, results that illustrate deficiencies in measures' ability to capture their intended construct. Nomological nets further suggested that the theoretical space of "racial prejudice" is crowded with scales that may not actually capture meaningfully distinct latent constructs. We provide concrete recommendations for scale selection and renovation and outline implications for overlooking measurement issues in the study of prejudice and discrimination.

149/250

Key words: stereotyping; prejudice; latent constructs; measurement; psychometrics

## Evaluating Validity Properties of 25 Race-Related Scales

Attitudes, beliefs, and motivations concerning race are central to many prominent theoretical perspectives on prejudice and discrimination. Accordingly, researchers have developed and used scales to measure the effect of race-related attitudes on a wide variety of outcomes. Yet the capacity for these theories to explain behavior hinges on how well researchers are accurately measuring these latent constructs. To measure a construct poorly is to introduce error, leaving one unable to test hypotheses with precision. Just as an old metal detector will undoubtedly find some rings and coins but leave other treasure undiscovered, so too will an outdated or poorly designed scale reveal some effects but also leave many others undiscovered or poorly estimated. Similarly, while an old metal detector might falsely signal the presence of gold when there are actually only iron oxides beneath the surface, the extent to which a scale fails to capture its intended construct will also lead researchers to draw erroneous conclusions about the theoretical meaning of observed effects.

A variety of scales have been developed and used by researchers to capture various facets of explicit racial attitudes, beliefs, and motivations. Approaches include: asking people directly about their level of racial prejudice (Axt, 2018), their race-related political attitudes (Henry & Sears, 2002), whether race contributes to the accuracy of various judgments (Uhlmann et al., 2010), whether they are motivated to control their own prejudice (Plant & Devine, 1998), their knowledge of cultural stereotypes (Ghavami & Peplau, 2013), how much conflict they perceive between groups in society (Sidanius et al., 2004), and their endorsement of racism-adjacent attitudes such as Right Wing Authoritarianism (Altemeyer, 1988) and Social Dominance Orientation (Pratto et al., 1994). Additionally, some scales were constructed to capture variation in attitudes toward Black people more generally, rather than to measure a specific race-related

attitude (e.g., the American National Election Survey scale; Payne et al., 2010). Notably, some scales "cluster" together such that they are related, sharing similar items, origins, or theoretical motivations, yet are still somewhat distinct. For example, the Symbolic Racism 2000, Modern Racism, and Racial Resentment Scales can all be understood as offspring of broader theorizing about Symbolic Racism (see Sears, 1988).

We broadly refer to this cluster of scales in the literature as "race-related scales", not because they were all designed to specifically capture racial attitudes, beliefs, and motivations, but because they are either functionally used for this purpose (see Axt, 2018 for discussion) or used to explain racism-related attitudes and outcomes (e.g., Right Wing Authoritarianism; see Duriez & Soenens, 2009; Hiel & Mervielde, 2005; Nicol & Rounding, 2013). The broad evaluation of these race-related scales includes many that have shown a marked influence on psychological research, with fourteen of the scales' papers amassing over 500 citations and four of the scales' papers amassing over 2500 citations (see Table 1). Furthermore, from a practical perspective, these scales are associated with race-related outcomes via their inclusion in Project Implicit data collection alongside measures of implicit racial prejudice. This dataset constitutes one of the richest and most influential sources of information on racial attitudes.

Measuring constructs as well as possible and with minimal error is key to hypothesis testing (Flake & Fried, 2020). Good measurement is not merely a concern for the replicability or reproducibility of results, but a key element of precisely connecting data to theory: if you do not know what you are measuring, or measuring it poorly, any results are dubious. Indeed, some scholars have argued that there is a "theory crisis" in psychology that partially stems from invalid measurement of latent constructs (Eronen & Bringmann, 2021). Any researcher hoping to tap racial attitudes, beliefs, or motivations must choose carefully between the numerous

measurement options, and it is difficult to holistically consider the multi-faceted evidence about the quality of many scales. Here, we address this important concern using a large dataset and modern methodology to evaluate the validity properties (i.e., construct validity in general with a greater focus on structural validity) of 25 race-related scales.

Our intention is not to show the invalidity of any given scale. Indeed, the scales that we evaluate have played essential roles in decades of research on the nature of racial stereotyping and discrimination. Instead, we aim to identify which scales currently have the best psychometric properties and highlight opportunities to renovate existing scales to better capture the underlying latent factors they are designed to measure.

**The Ongoing Process of Construct Validation**

Loevinger (1957) categorized the process of construct validation into three phases: substantive, structural, and external. The substantive phase outlines the theoretical underpinnings of a construct. The structural phase involves quantitative analyses, examining the psychometric properties of the measure like reliabilities and factor structure. Finally, the external phase measures whether the scale relates to the attitudes and outcomes one would expect it to predict, such as other measures of similar constructs as well as relevant judgments and behaviors.

Researchers who develop and use scales often overlook or downplay the structural phase of construct validation. For instance, in a broad examination of construct validation in social and personality psychology, 57 of 301 reviewed scales provided no information at all about the scale, and another 205 provided only information about internal reliability (largely via Cronbach's α; Flake et al., 2017). In our more targeted review of the race-related scales evaluated in this

manuscript, we were only able to locate clear reliability information for 20 of the 25 scales (see Table 1).

Table 1

*Information about the 25 race-related scales*.

| Scale Name | Citation | Number of Citations | Number of Items | Sample Size | Population Sampled | Internal Reliability Test | Factor Structure Test |
|---|---|---|---|---|---|---|---|
| American National Election Survey | (Payne et al., 2010) | 244 | 6 | 1933 | Representative American sample (with oversampling of Black and Latine people) | Cronbach's α | n/a |
| Attitudes Toward Blacks | (Brigham, 1993) | 545 | 20 | 260 | Undergraduate students | Cronbach's α | Principal component analysis |
| Attitudes Toward Whites | (Brigham, 1993) | 545 | 20 | 81 | Undergraduate students | Cronbach's α | Principal component analysis |
| Bayesian Racism | (Uhlmann et al., 2010) | 60 | 6 | 109 | American adults | Cronbach's α | n/a |
| Cultural Attitudes Toward Black People | (Nosek & Hansen, 2008) | 224 | 6 | >100k | Students and Project Implicit respondents | Cronbach's α | n/a |
| Cultural Attitudes Toward White People | (Nosek & Hansen, 2008) | 224 | 6 | >100k | Students and Project Implicit respondents | Cronbach's α | n/a |
| General Intergroup Anxiety | (Stephan et al., 1999) | 812 | 12 | 332 | Students from Florida, New Mexico, Hawaii | Cronbach's α | Principal component analysis |
| General Social Survey, Race Items | (Davis & Smith, 1991) | 255 | 22 | n/a | American adults | n/a | n/a |
| Intergroup Anxiety | (Britt et al., 1996) | 190 | 11 | 2551 | Students from Kansas and Florida | Cronbach's α | n/a |
| Internal and External Motivation to Control Prejudice | (Plant & Devine, 1998) | 1860 | 10 | 1743 | White psychology students | Cronbach's α | Principal component analysis and confirmatory factor analysis |
| Motivation to Control Prejudiced Responses | (Dunton & Fazio, 1997) | 1044 | 17 | 1109 | College students | Cronbach's α | Principal component analysis |
| Modern Racism | (McConahay, 1983) | 589 | 7 | 81 | White Duke University students | Cronbach's α | n/a |
| New Racism | (Jacobson, 1985) | 307 | 7 | 1429 | White adults | Cronbach's α | Principal component analysis |

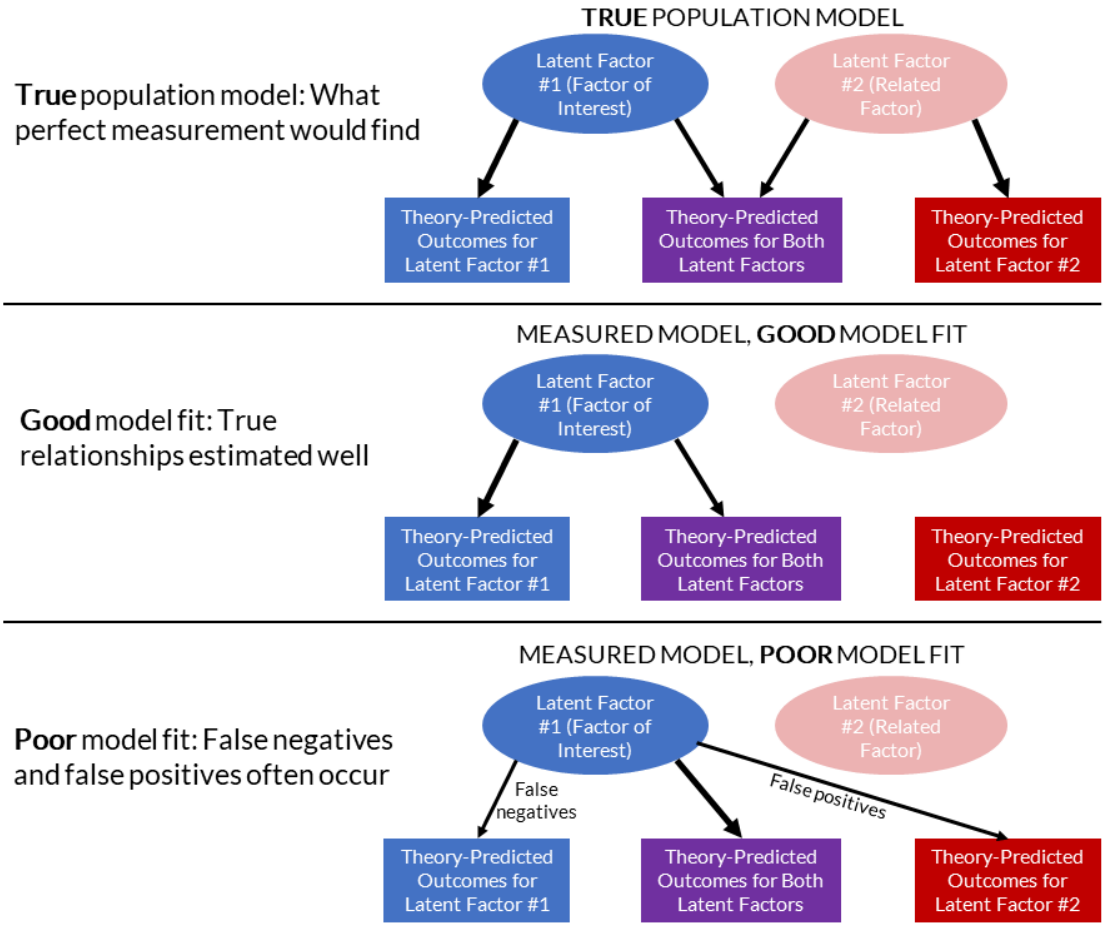
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Pro-/Anti- Black Attitudes Questionnaire | (Katz & Hass, 1989) | 1825 | 20 | 1104 | Students from eight universities | Cronbach's α | Principal component analysis |
| Perceived Group Conflict | (Sidanius et al., 2004) | 309 | 6 | 2132 | UCLA students | Cronbach's α | n/a |
| Prejudice Index | (Bobo & Kluegel, 1993) | 1031 | 5 | 1309 | American adults from the 1990 General Social Survey | n/a | n/a |
| Perceptions of Others' Preference | n/a | n/a | 6 | n/a | n/a | n/a | n/a |
| Racial Attitudes | (Sidanius et al., 1991) | 217 | 14 | 5655 | University of Texas students | Cronbach's α | Principal component analysis |
| Racial Arguments | (Saucier & Miller, 2003) | 90 | 13 | 942 | White students | Cronbach's α | Principal component analysis |
| Racial Resentment | (Kinder et al., 1996) | 2622 | 6 | n/a | White American adults from 1986 National Election Study | Cronbach's α | n/a |
| Racial Stereotypes Measure | (Peffley et al., 1997) | 501 | 5 | 1841 | White American adults from 1991 Race and Politics Survey | Cronbach's α | n/a |
| Right-Wing Authoritarianism | (Altemeyer, 1988) | 3025 | 20 | n/a | n/a | n/a | n/a |
| Subtle and Blatant Prejudice Scale | (Pettigrew & Meertens, 1995) | 2618 | 20 | 3810 | Adults in France, the Netherlands, Great Britain and West Germany. | Cronbach's α | Principal component analysis |
| Social Dominance Orientation | (Pratto et al., 1994) | 5190 | 16 | 1952 | Stanford students | Cronbach's α | n/a |
| Symbolic Racism 2000 | (Henry & Sears, 2002) | 998 | 8 | 887 | White adults in LA county; White UCLA students | Cronbach's α | Principal component analysis |

*Note.* A more detailed table is provided on the OSF page. Wherever possible, we use the original scale development paper. In cases in which no such paper is available (e.g., government-distributed scales), we cite a prominent paper using the scale and reporting metrics. The Perceptions of Others' Prejudice scale has no citation, but is available in the Project Implicit data.

Downplaying structural validity can appear innocuous when the substantive and external stages of construct validity appear to yield good evidence of a scale's functionality. However, a scale with good substantive and external validity can still lead to incorrect conclusions about the nature of the latent construct. For example, consider Figure 1, which depicts a hypothetical "true" model of two distinct but related factors (Factor 1 and Factor 2). With perfect measurement of both factors (top panel), results only find evidence that each factor predicts outcomes for which it is truly related. The same is true for when only one of the two factors is measured well (middle panel). However, with poor measurement fit, the substantive and external phases could yield good evidence. But since the scale is now capturing both factors and doing so with considerable error, accurate conclusions are jeopardized through higher rates of Type I (i.e., incorrectly concluding that a factor predicts an outcome it does not) and Type II errors (i.e., incorrectly concluding a factor does not predict an outcome that it does).

**Figure 1**

*Illustration of the Importance of Structural Validity*




*Note.* Arrow width represents the size of the relation between variables. In color figure, variables related to Factor #1 are blue; variables related to Factor #2 are red; and variables related to both factors are purple.

Even when these race-related scales included more rigorous evaluations of structural validity, the passage of time since their creation still poses a threat to scale validity. Construct validation is an ongoing process (Cronbach & Meehl, 1955), and how much information a given

scale provides about the underlying latent construct is context-dependent. Some scales that may have been highly reliable and valid in past decades may no longer be so due to cultural changes in society that have rendered their items less informative (Fabrigar & Wegener, 2016; Kane, 2013). For example, the question “Interracial marriage should be discouraged to avoid the 'who-am-I?' confusion that the children feel” from the Attitudes Toward Blacks scale (Brigham, 1993) might be interpreted differently three decades later. Furthermore, modern research now has a greater focus on how findings might be constrained or generalize beyond the sample on which it was developed (Henrich et al., 2010). Six of the 25 race-related scales we evaluated were validated only for White participants and 12 of the 25 scales were validated only for college students (see Table 1), and thus may not be suitable for capturing the attitudes of non-White participants.

**Modern Developments in Evaluating Structural Validity**

The ongoing process of construct validity does not just concern the shifting meaning of items and populations of interest. The specific tools used to evaluate aspects of validity and reliability have improved considerably in the past few decades. Furthermore, existing but underused tools have become very accessible thanks to advances in open-sourced statistical software. We incorporate four new or underused tools in social psychological measurement in the present work: McDonald’s ω to evaluate global internal consistency; dynamic fit indices to better evaluate model fit in confirmatory factor analysis; Item Response Theory to evaluate the distribution of latent factors and local reliability of scales; and nomological nets to generally evaluate the convergent and discriminant validity of scales by considering each scale’s relation to all other scales. None of these valuable modern tools were used for the validation of the 25 race-related scales that we review (see Table 1).

In the following sections, we discuss each of these tools, contrasting them with traditional methods when appropriate, and highlighting their advantages and unique contributions. We also describe the corresponding data analysis plan for evaluating the 25 race-related scales considered in this paper.

*McDonald's ω*

Internal reliability refers to the extent that the items in a scale are consistent with one another. Cronbach's α is the most-commonly used measure of global internal reliability (Cronbach & Meehl, 1955). Researchers typically only report coefficient α as a measure of internal consistency in social psychology (73%; Flake et al., 2017). However, Cronbach's α relies on a handful of assumptions that are rarely if ever met, such as complete unidimensionality and essential tau-equivalence (i.e., the equal loading of all items onto the latent factor; Dunn et al., 2014; Hayes & Coutts, 2020). The violation of these assumptions can bias Cronbach's α to overstate reliability. For this reason, researchers have encouraged the use of McDonald's ω as a more accurate measure of global internal reliability (McDonald, 2013), as McDonald's ω eschews assumptions of unidimensionality and essential tau-equivalence. We compare Cronbach's α and McDonald's ω for all scales.

*Dynamic Fit Indices*

Although measures of internal consistency such as Cronbach's α and McDonald's ω are related to evaluations of factor structure such as confirmatory factor analysis (CFA), they are not equivalent. In CFA, researchers impose a model on the data in which one or more underlying latent factors are theorized to "cause" the responses to the items in a survey. Various model fit indices, such as the Comparative Fit Index, Root Mean Square Error of Approximation, and Standardized Root Mean Square Residual, attempt to capture the model's alignment with the

data, allowing researchers to make informed decisions about whether a model is a "good" fit for the data. Poor model fit indicates some imprecision about the structure of the data, which essentially translates to not measuring the construct that one thinks one is measuring. If the poor-fitting scale is then used to predict outcomes, any conclusions rendered using this scale are more likely to be wrong, due to uncertainty about what exactly the scale is really measuring.

While creating guidelines allows for a wider adoption of these methods, exactly what constitutes "good" or "bad" model fit can be unclear. In a seminal paper, Hu and Bentler (1999) provided "rule-of-thumb" model fit thresholds against which researchers could evaluate their models. Researchers took full advantage of these concrete guidelines—the paper now has over 73000 citations. However, the model fit thresholds defined by Hu and Bentler are based on models with very specific characteristics on many dimensions, such as factor loadings, number of latent factors, and correlation between latent factors. For example, the "reliability paradox" describes how a scale with less measurement error can actually have *worse* model fit than a scale with high measurement error, even if they appear to have the same model fit statistics (Hancock & Mueller, 2011; McNeish et al., 2018). Although Hu and Bentler warned against blanket use of their model fit thresholds across all CFA contexts (Hu & Bentler, 1998, p. 446; see Barrett, 2007; Millsap, 2007), researchers have typically applied them without considering the constraints of the original simulations.

Dynamic fit indices address this shortcoming in evaluating model fit (McNeish & Wolf, 2021). This approach conducts simulations to generate appropriate model fit thresholds on a model-by-model basis, taking into account specific characteristics of the model including: loadings, item intercepts, number of items, sample size, error variance, number of latent factors, and the correlation between latent factors. Doing so generates model-specific fit thresholds that

match the intended use of the static model fit thresholds to effectively balance Type I and Type II errors when evaluating factor structure. We evaluate each race-related scale using both Hu and Bentler cutoffs and dynamic fit indices to contrast these results.

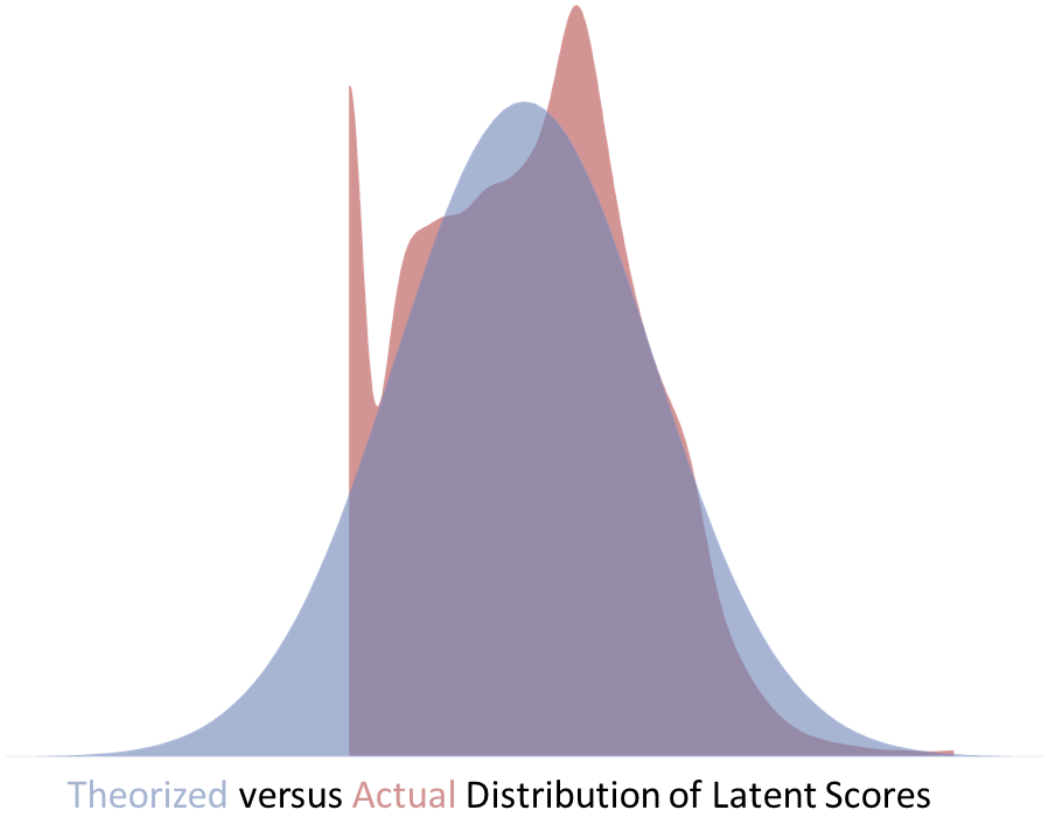
*Item Response Theory: Local Reliability and the Distribution of Latent Scores*

Evaluations of construct validity go beyond factor structure. Construct validity also concerns whether the theorized distribution of latent factor scores (i.e., the expected distribution of the latent factor in the population) can be adequately captured by the scale items. Imagine a racial prejudice scale that assumes a normal distribution of racial prejudice across the population (Figure 2). The latent scores show that the lowest score on the scale is also the mode, far from the theorized normal distribution. This pattern would suggest that the scale is not adequately capturing the theorized distribution of the latent factor in the population due to floor effects. As a result of this floor effect, any researcher interested in capturing sample variation at the lower end of the latent construct would be unable to do so. For example, imagine researchers were interested in correlating prejudice scores with an outcome within a population low in anti-Black prejudice (e.g., students at historically Black colleges and universities). If these researchers used a scale with a floor effect, they would erroneously find very little variation in anti-Black prejudice and likely to draw incorrect conclusions from their data, because the scale does a poor job separating those low in prejudice from those very low in prejudice.

Item Response Theory (IRT) can provide insight into the distribution of latent scores. After fitting a model, latent factor scores are predicted for each individual in the sample and plotted in a distribution to identify potential levels of the latent factor not well-captured by the scale.

**Figure 2**

*Theorized versus actual latent score distributions*




*Note.* These hypothetical data illustrate how the theorized distribution of latent factor scores may not be matched by the actual distribution of the measured latent factor scores.

Of course, the idea that internal reliability is global (i.e., stable across levels of the latent factor) is itself a major assumption that typically goes unexamined in social psychology. The internal reliability of a scale can vary as a function of the mean (i.e., high to low) of the latent factor, and this localized variation in reliability can be examined using IRT (Baker, 2001). For example, the Need for Cognition scale shows high internal reliability overall, but reliability decreases at the positive end of the scale (i.e., for people high in Need for Cognition; Edwards, 2009). As another example, the Affect Scale (Zanon et al., 2013) shows better reliability at lower

levels of positive affect (Zanon et al., 2016). For researchers targeting populations with lower or higher levels of racial attitudes, considering localized internal reliability is critical. Here, we used IRT to examine local reliability for all scales.

*Nomological Nets*

Finally, we consider the relationship between the 25 race-related scales by constructing a nomological net. This nomological net allows us to broadly evaluate both the convergent and discriminant validity of each scale (Cronbach & Meehl, 1955). It is important to note that a nomological net doesn't tell you about *what* you are measuring, only the extent to which any two scales are correlated. Yet if two scales are located very closely in a nomological net, and one purports to measure construct X, and another to measure a distinct construct Y, we can correctly infer that one of these claims is likely incorrect. Highly correlated scales located in a similar space suggests that the latent constructs measured may be similar, even if they purport to measure something distinct. Furthermore, the nomological net provides more global information about which areas of the latent factor "space" are more densely populated with scales. Similarly, this latent factor space also identifies areas that are sparsely populated, highlighting scales that are capturing something unique.

**The Present Study**

In this study, we evaluated the validity properties of 25 race-related scales. We used a Project Implicit dataset (Axt, 2018) with over one million participants completing 2 of the 25 scales—the dataset contains over 40,000 responses to each of the 25 scales. This evaluation is the most thorough to date, featuring sample sizes at least 16 times larger than those used to validate the scales in the original papers. Additionally, despite the limited and non-representative nature of the Project Implicit sample, the sample is still far more representative than the samples

used in the initial validation of nearly all of these scales (see Table 1 for comparison). Compared to original works that overwhelmingly used (mostly White) undergraduates in psychology classes, the present sample is larger, and with greater racial, ethnic, and age diversity. Finally, because the Project Implicit dataset includes all 25 scales in the same sample, it is uniquely suited to creating a nomological net of these scales, a key aspect of establishing construct validity (Cronbach & Meehl, 1955; Flake et al., 2017). To our knowledge, this is the first time such a broad network in the prejudice domain has been established. However, one concern about a Project Implicit sample is its representativeness, given that participants self-selected into the study. To better generalize our findings, we also examined a smaller supplementary sample of paid, online participants.

## Method

All data, scripts, figures, analysis markdowns, and other supplementary materials are available at https://osf.io/zg6fr/?view_only=e6d56172a5934259a81729312ebf0754. Markdowns are recommended as the most accessible way to evaluate details of the methodology. We did not complete preregistrations for this project.

### Participants

We used data provided by 1,396,234 Project Implicit respondents (60.1% Female, 68.5% White, 9.7% Black, $M_{age}$ = 27.3 years, $SD_{age}$ = 12.2, 82.8% US residents) between October 23, 2014 and September 27, 2016. This data was originally analyzed in Axt (2018). Each respondent was asked to complete a demographics questionnaire, the Race Implicit Association Test, as well as two randomly-selected race-related scales (to avoid participant fatigue). Order of measures was randomized to account for any possible order effects. 365,027 participants dropped out of the study before completing the explicit race-related scales. We also excluded respondents

outside of North America (US and Canada), given the unique racial context in North America that many of these scales are originally intended to capture (analyses including these respondents are available on the OSF page). With these exclusions in mind, we conducted analyses on datasets including both White respondents only ($N$ = 569,414) and all respondents ($N$ = 910,066).[1] We center the analyses including only White North Americans in our figures and reporting.

## Materials

We evaluated 25 scales in this study,[2] and thus refrain from providing an in-depth description of each scale. See Table 1 for information about each scale.[3] The wording for each individual scale item is provided on the OSF page, as are any deviations from the original wording.

## Analytic Approaches

Analyses were completed in R using *lavaan* (Rosseel, 2012) for CFA, *dynamic* to generate dynamic fit indices for CFA (McNeish & Wolf, 2021), *ltm* (Rizopoulos, 2006) for IRT, and *igraph* (Csardi & Nepusz, 2006) for creating the nomological network.

### *Alpha and Omega*

Although CFA can provide more in-depth information about internal consistency, Cronbach's α and McDonald's ω are still commonly reported metrics throughout the literature. Further, we believe our analyses use the largest validation sample size to date for all scales. Accordingly, we considered it valuable to report and compare how these scales performed on each of these metrics, enabling easy contrasts by researchers in their own work. Here, to give a

[1] All respondents included respondents that did not report their race.
[2] Some sections refer to "30 scales and subscales", because 5 of the scales are comprised of two separate factors.
[3] Note that the results for Attitudes Toward Whites are omitted from the figures for internal consistency and model fit, as this scale performed substantially worse than all other scales.

more direct comparison of how Cronbach's α and McDonald's ω compare to each other, we use McDonald's $\omega_u$, which assumes that the latent factor is unidimensional and that the indicators are continuous (Flora, 2020).

*Confirmatory Factor Analysis*

**Evaluations of model fit.** We focused on three commonly reported indices: the Standard Root Mean Squared Residual (SRMR), the Root Mean Square Error of Approximation (RMSEA), and the Comparative Fit Index (CFI). For each of these statistics, we compared the actual fit to both the commonly used Hu and Bentler (1999) rule-of-thumb thresholds and the dynamic fit thresholds. Dynamic fit thresholds were calculated as a function of model factor loadings, item intercepts, number of items, sample size, error variance, number of latent factors, and the correlation between latent factors. These thresholds correctly reject misspecified models 95% of the time, while incorrectly rejecting correctly-specified models 5% of the time. For full details, see (McNeish & Wolf, 2021).

**Methods of estimation.** For our CFAs, we used two different estimation approaches. First, we used Maximum Likelihood, traditionally used for the estimation of latent constructs in social and personality psychology. Currently, the estimation of dynamic fit indices is only fully compatible with Maximum Likelihood models and other models that treat latent factor indicators as interval, making the use of Maximum Likelihood necessary for the estimation of dynamic model fit thresholds.

Second, we used robust Diagonal Weighted Least Squares estimator, which treats the latent factor indicators as ordinal instead of interval (the latent factor itself is still on an interval scale). Diagonal Weighted Least Squares provides more unbiased factor loadings and fit statistics for scale items under most conditions (Li, 2016a, 2016b). Further, because Likert-type

items are more accurately characterized as ordinal rather than interval, they better represent the data. Although dynamic fit indices are not currently designed for use with Diagonal Weighted Least Squares models, we nevertheless considered both dynamic thresholds and Hu and Bentler thresholds to comprehensively evaluate these scales. These results are provided on the OSF page.

*Item Response Theory: Latent Factor Distribution and Local Reliability*

We used IRT to evaluate the distribution of latent factor scores alongside the local reliability for each race-related scale, with an emphasis on evaluating those that show reasonably good model fit and/or are prominent in the social and personality psychology literature. Latent factor distribution is essentially how well the latent factor is measuring a construct across different levels of the scale, which we examined by predicting latent factor scores using IRT Models and then plotting a density function for these latent factor scores. Local reliability is captured by the Test Information Function, which illustrates the amount of information provided by the test items across levels of the latent distribution (Baker, 2001; Edwards, 2009). For interpretability, we used the formula $\sqrt{1/INFORMATION}$ to convert Information to the Standard Error of Measurement,[4] which describes the extent to which an observed score likely differs from the true score (Dudek, 1979; Edwards, 2009; Tighe et al., 2010).

Latent factor distribution and local reliability are related. Scales that show steep declines in reliability at values of a latent factor also tend to show "peaks" indicating floor or ceiling effects in the scale's ability to measure the latent factor. For this set of analyses, we focused primarily on latent factor distributions. "Peaks" in the observed distribution at the edges of the scale indicate ceiling or floor effects, typically accompanied by local reliability issues at the scale extremes.

[4] Distinct from standard error.

Complete output from the IRT analyses are available on the OSF page. They also provide discrimination parameters for each item in each scale, as well as the graded-response model extremity scores for the outermost responses to each scale item.

*Nomological Networks*

Nomological networks are a representation of constructs and the relationships between them (Cronbach & Meehl, 1955). These nets help assess whether a construct is "behaving as theorized" within a broader constellation of other constructs. In other words, it should be closer in space to other similar constructs, and further from those theoretically posited to be dissimilar. The distance between concepts is a function of some measure, such as the correlation between any two constructs.

In the present research, we created nomological nets using the Pearson correlation between Diagonal Weighted Least Squares latent scores. A force-directed algorithm determined the positioning of all the constructs relative to one another (Kamada & Kawai, 1989).

## Results

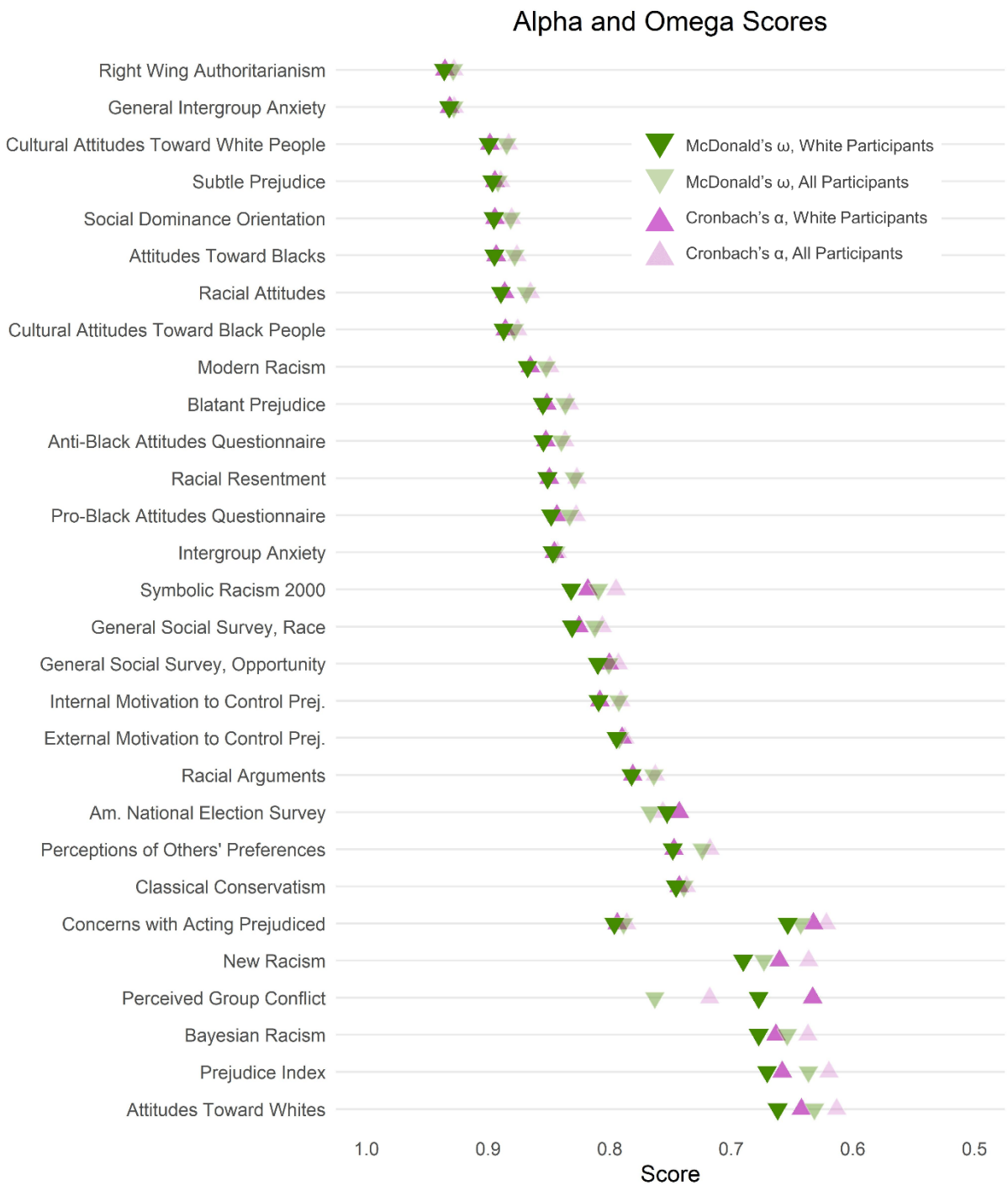
### Alpha and Omega

Cronbach's α and McDonald's ω equations yielded similar internal reliability scores, with some exceptions. Most scales, but not all, showed adequate global internal reliability by commonly-used standards, with 25 of the 30 scales and subscales showing both Cronbach's α and McDonald's ω values of over .70. Reliability for many of these scales is higher for White participants, compared to all participants, consistent with the development of many of these scales to measure the attitudes of White people (Figure 3). See the OSF page for full tables listing Cronbach's α and McDonald's ω scores.

When Cronbach's α and McDonald's ω do diverge, it is likely because of Cronbach's α assumption that the item variances of the true scores are constant across items (a *tau-equivalent model*) has been violated. McDonald's ω makes no such assumption, allowing the item variances of the true scores to differ from item to item (a *congeneric model*, which is more consistent with CFA; see Dunn et al., 2014).

**Figure 3**

*Cronbach's α and McDonald's ω scores*




*Note*. Scales are arranged such that scores progress from left (best score) to right (worst score). This practice is maintained throughout the paper.

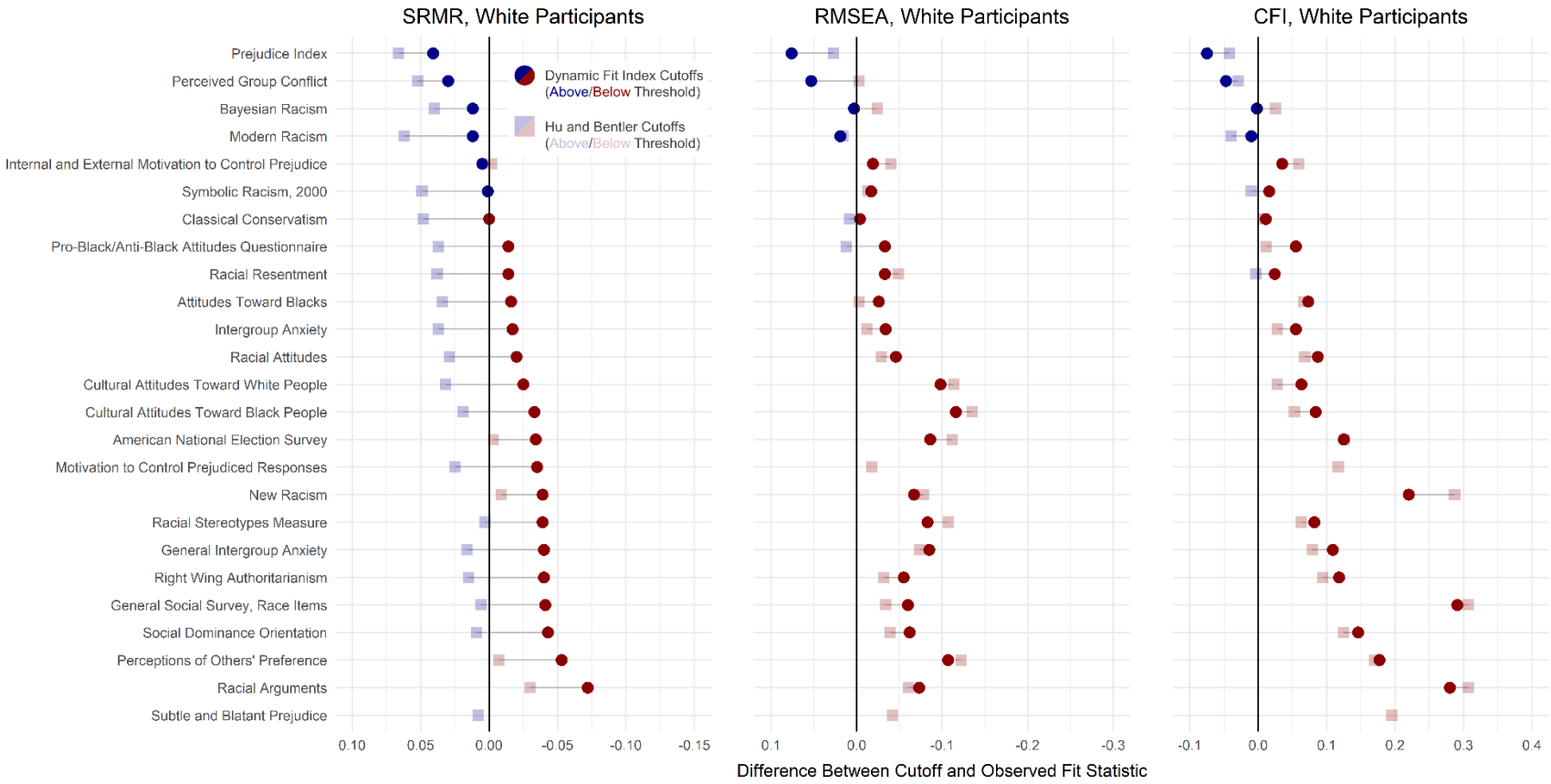
**Confirmatory Factor Analysis**

In this section, we considered the extent to which Maximum Likelihood CFA models for each of the scales adequately fit the observed data. For each of these models, we calculated the difference score between the observed SRMR, RMSEA, and CFI fit statistics and the dynamic fit cutoff produced using McNeish and Wolf's (in press) methodology. We also calculated the difference score between the observed SRMR, RMSEA, and CFI fit statistics and the Hu & Bentler (1999) traditional cutoffs, to illustrate the difference between the traditional and dynamic cutoffs. We examined the model fit of entire scales rather than individual subscales. For every scale with multiple factors, we fit the models theoretically proposed by the authors.

Results were similar for all participants and White participants only, and were also similar when comparing Maximum Likelihood Results to Diagonal Weighted Least Squares results. See Figure 4 for fit statistics for Maximum Likelihood using all participants. See the OSF page for fit statistics from Maximum Likelihood using White participants only and all Diagonal Weighted Least Squares models.

**Figure 4**

*Model Fit Information for Both Traditional and Dynamic Cutoffs*




*Note.* Values indicate difference scores subtracting the fit statistic from the cutoff statistic. Scales are arranged such that the best SRMR fit statistics (to the left) are at the top and the worst SRMR fit statistics (to the right) are on the bottom.

For the majority of scales, the theoretical latent factor poorly fit the observed data. In the case of SRMR, whereas 18 of the 25 scales pass the traditional Hu and Bentler cutoff, only 5 of the 25 scales pass the dynamic fit cutoff necessary to correctly rejected misspecified models 95% of the time. On the other hand, differences between the two types of cutoffs were more modest for RMSEA and CFI, with certain scales (most noticeably Prejudice Index and Bayesian Racism) showing *better* evidence of good model fit with the dynamic cutoffs than with the Hu and Bentler cutoffs. Given that the evaluation of model fit is typically performed taking all of these indices into consideration (e.g., Hussey & Hughes, 2020), the differences in results between the dynamic and traditional cutoffs change the conclusions researchers might reach about the model fit of a given scale. Overall, these dynamic cutoffs illustrate that many scales commonly used by social psychologists are likely misspecified to some degree and that some scales show poor enough model fit that they include a substantial amount of error.

**Distribution of Latent Scores and Local Reliability**

Here, we evaluated to what extent a particular scale is more or less likely to capture attitudes at certain values of the scale. In Figure 5, we visualize these results for six race-related scales, selected either for their good dynamic model fit (Bayesian Racism, Modern Racism, Perceived Group Conflict, Prejudice Index) or for their importance as theoretical constructs in the literature (Racial Resentment, Social Dominance Orientation). The density plots depict the distribution of latent factor scores and the lines represent the Standard Error of Measurement as a function of latent factor level. Graphs are available for all 25 scales on the OSF page.

Of note, Modern Racism and Perceived Group Conflict show severe floor effects, such that the scale fails to distinguish between individuals low in these latent factors. Bayesian Racism and Social Dominance Orientation show modest floor effects, and Racial Resentment

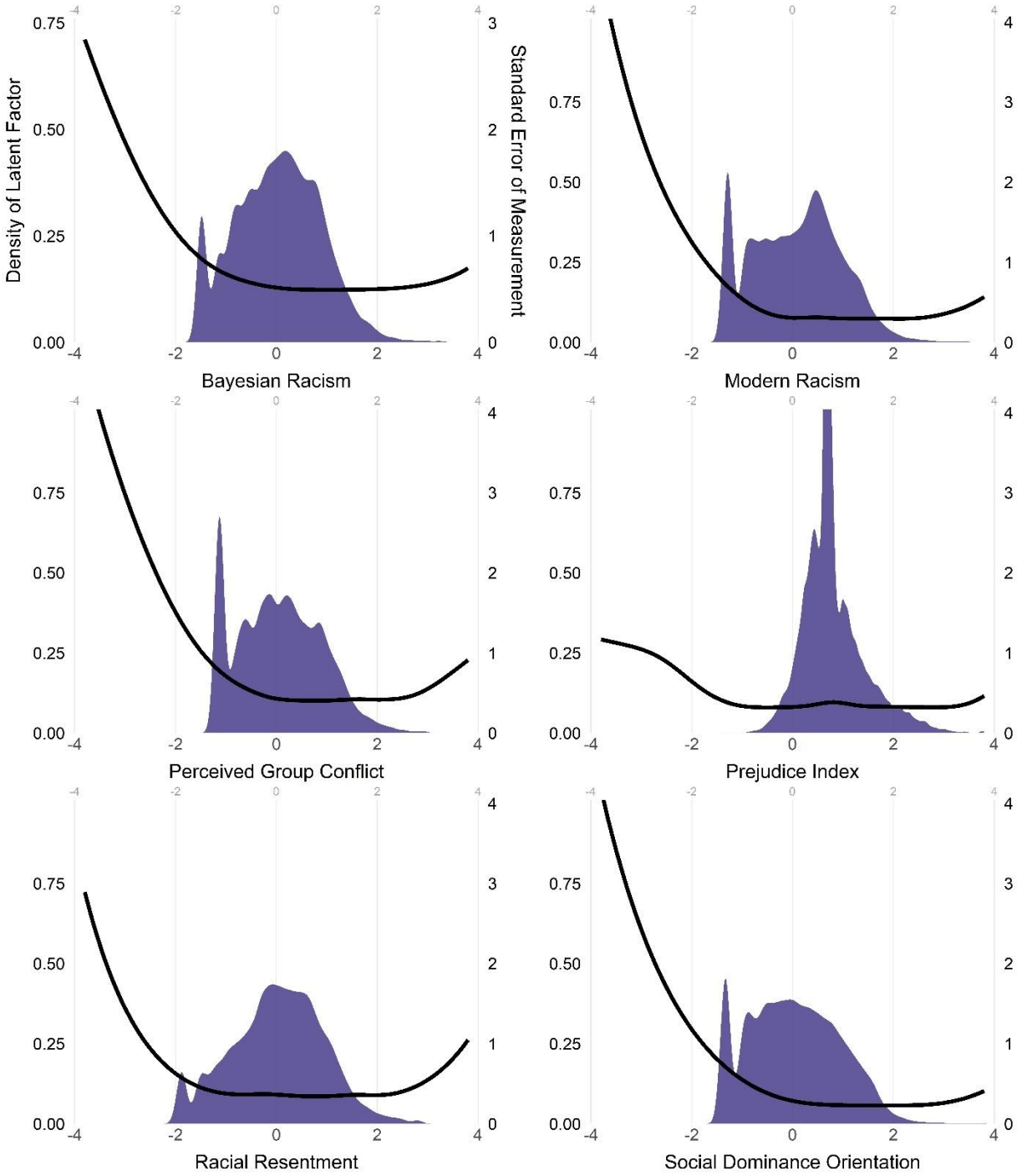
shows a minor floor effect but follows a relatively normal distribution. Notably, a severe increase in the Standard Error of Measurement accompanies these peaks, indicating low reliability of the scale items for participants whose “true score” on the latent factor is low.

The Prejudice Index shows a very large peak at the center of the distribution, consistent with the measure’s scoring being derived from a series of difference scores concerning the degree to which Black vs. White people have certain characteristics. This is less problematic than a concentration of scores at the edge of the distribution. The scale appears to distinguish between strong pro-Black attitudes, neutral attitudes, and strong pro-White attitudes, unlike the rest of the scales (perhaps excluding Racial Resentment). This interpretation is consistent with the stability of the Standard Error of Measurement toward the center of the distribution.

Results were similar for analyses restricted to White participants. An examination of the other nineteen scales shows floor effects for both General Intergroup Anxiety and Intergroup Anxiety and a large ceiling effect for Internal Motivation to Control Prejudice. See the OSF page for information regarding item-level discrimination and extremity parameters for all scales, which is useful for closely examining the contents of a specific scale.

**Figure 5**

*IRT Latent Factor Distributions and Standard Errors of Measurement*



*Note.* The purple density plot depicts the distribution of predicted latent factor scores for each participant. The black line indicates the Standard Error of Measurement. The x-axis unit is the standardized latent factor in IRT (i.e., Theta).

**Nomological Net**

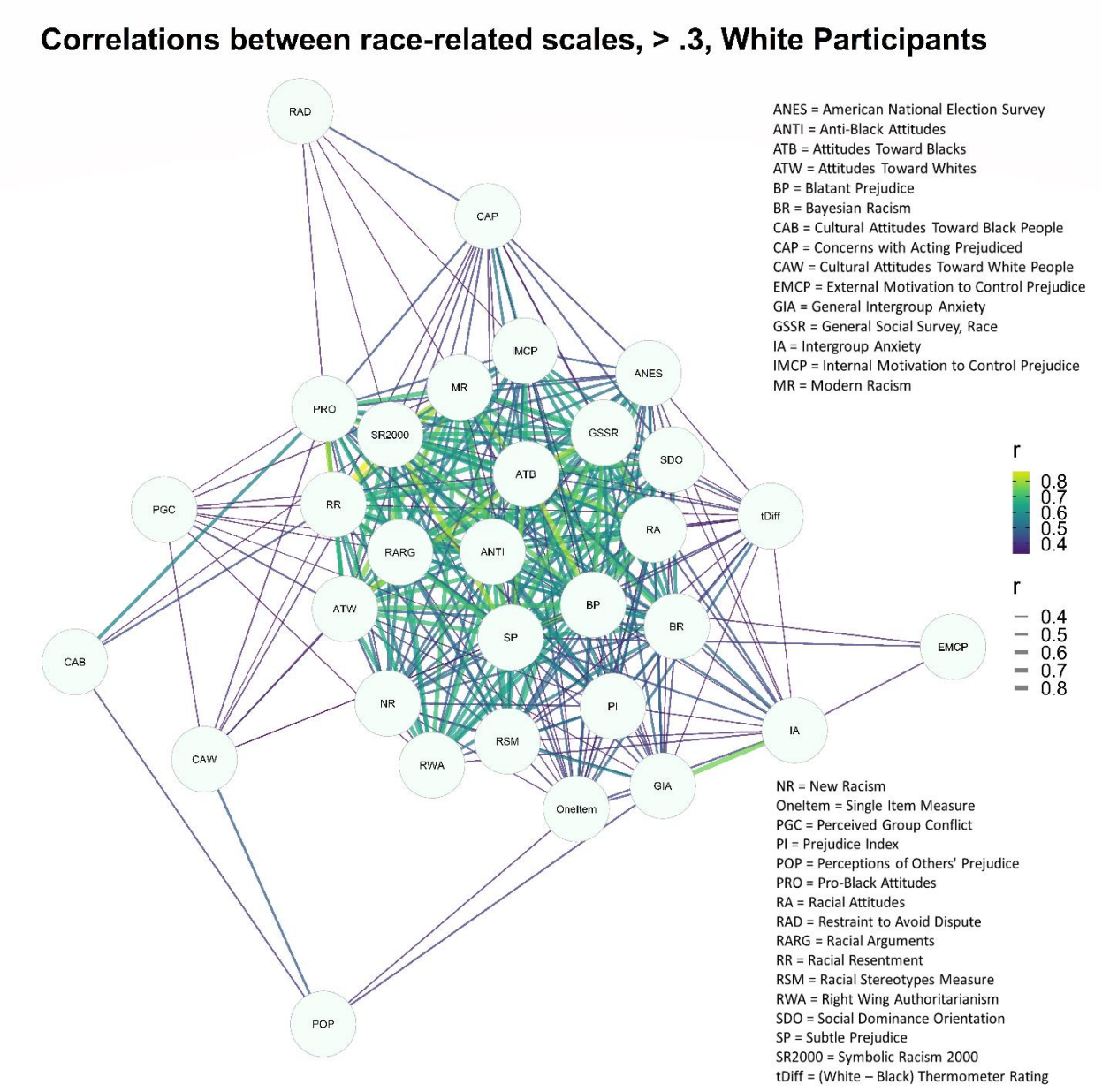
We created a nomological net featuring all evaluated scales and subscales (30 total) using latent factor scores (Figure 6). The most noticeable feature of this net is the tight clustering of the majority of the scales. This is consistent with an interpretation that these scales are all tapping similar and related constructs, even when designed to measure attitudes, motivations, or beliefs about groups in general rather than racial groups specifically. What those constructs are, exactly, cannot be determined from this analysis, but many are theorized to measure racial prejudice, though the cluster also includes some scales that are not theorized to directly measure racial prejudice (e.g., Internal Motivation to Control Prejudice, General Intergroup Anxiety, Intergroup Anxiety), scales theorized to be independent personality constructs (e.g., Social Dominance Orientation, Right Wing Authoritarianism), and scales purposely constructed to tap variability in attitudes toward Black people in a variety of domains (e.g., American National Election Survey). What this means is that, even if a scale was not designed to measure prejudice *per se* but is highly correlated with another designed to measure prejudice, it might be the case that at least one of the scales is being misinterpreted. Both may be tapping prejudice, or both may be tapping something else.

Notably, the two most straightforward measures of prejudice—a single 7-point measure of preference for White versus Black individuals (“OneItem”) and a difference score between 10-point thermometer ratings of White and Black individuals (“tDiff”)—are on the edge of the

central cluster and are less strongly related to many of the other race-related scales, though they are more strongly correlated with implicit attitudes than the other scales (Axt, 2018). The network also illustrates that certain scales occupy less-populated theoretical spaces. While we cannot know what, exactly, these scales are capturing, this visualization makes clear the relative sameness or distinctiveness of each scale.

**Figure 6**

*Nomological network of scales from White participants*




*Note.* Line width and line color are both functions of the strength of the correlation. Only relationships above .3 are plotted. The proximity of nodes reflects the relative position of each scale given its correlation with all other scales.

Measures of motivation to control prejudice (with the exception of External Motivation to Control Prejudice) and cultural knowledge of stereotypes occupy the area outside the main cluster, suggesting their relative distinctiveness as latent constructs. Finally, we note that Perceived Group Conflict is barely to any of the other scales in the nomological net, in line with its intent to capture experiences of discrimination, rather than prejudiced attitudes (Sidanius et al., 2004). For researchers interested in more closely examining the connections in the dense central cluster, a nomological net depicting only correlations of .5 and above is available on the OSF page.

## Robustness Checks

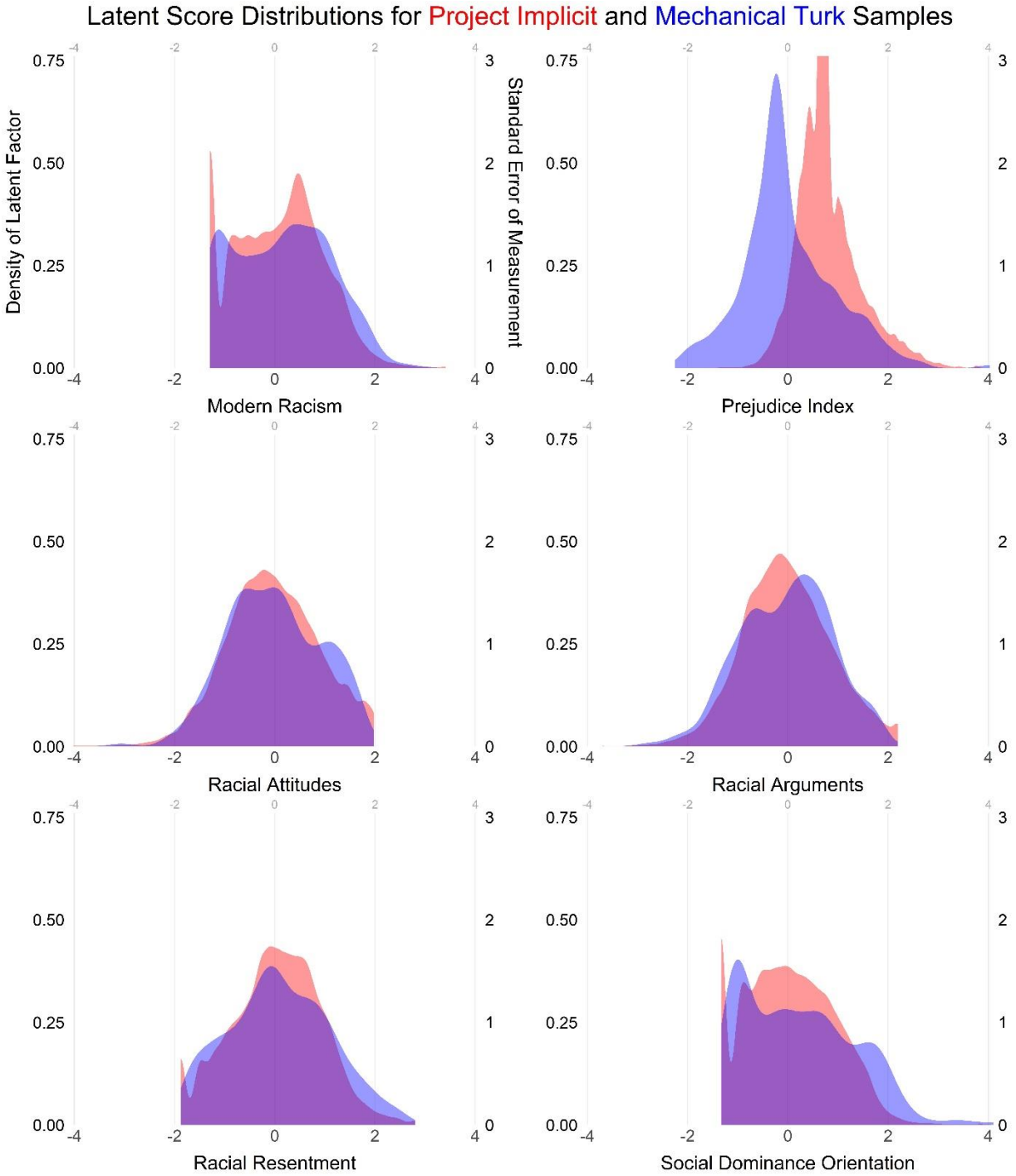
Despite being the largest and most representative samples to examine the majority of these scales, one might be concerned that these results do not generalize beyond the volunteer Project Implicit sample. In particular, we considered the possibility that the floor and ceiling observed in the latent score distributions might be a characteristic of the Project Implicit sample, who self-selected into the study and as a result may be more concerned about appearing unprejudiced. To explore this issue, we collected two separate samples from Mechanical Turk, an extremely common source of participants in modern psychology. For both theoretical and practical reasons, we opted to focus on the distribution of latent scores, which can easily and clearly be compared across samples despite that large difference in sample size. However, alpha, omega, and model fit statistics are available on the OSF page.

To facilitate useful comparisons between the latent score distributions in the two samples, we selected scales with relatively good, moderate, or bad fit in the Project Implicit sample that were further characterized by either distinctive (i.e., large floor effects, ceiling effects, or central peaks) or relatively normal distributions. In the first sample ($N$ = 308, $N_{White}$ = 280, 42.9% men,

56.8% women, .3% nonbinary, $M_{age}$ = 42.5 years, $SD_{age}$ = 14.2 years), participants provided responses for Modern Racism and Prejudice Index (good fit but non-normal latent score distributions). In the second sample ($N$ = 300, $N_{White}$ = 232, 53.8% men, 44.1% women, .7% nonbinary, $M_{age}$ = 40.8 years, $SD_{age}$ = 13.4 years), participants provided responses for Racial Resentment, Racial Attitudes, Racial Arguments, and Social Dominance Orientation (moderate to bad fit with various latent score distributions). Both samples were collected in 2021.

The IRT latent score distributions for White North American participants (both Project Implicit and Mechanical Turk samples) are depicted in Figure 7. Overall, results were extremely similar between the two samples. As in the Project Implicit sample, the distribution of latent scores in the Mechanical Turk sample showed floor effects for both the Modern Racism and Social Dominance Orientation scales. The Prejudice Index still demonstrates a noticeable peak in latent scores close to the center of the distribution. Finally, the other three scales' distributions resembled those observed in the Project Implicit sample, with the exception of a slight floor effect for Racial Resentment that is present in the Project Implicit data but reduced in severity in the Mechanical Turk data. We interpret results as evidence that results of the present research are not merely a function of the Project Implicit sample (or the Mechanical Turk sample).

**Figure 7**




*Note.* Project Implicit (red) and Mechanical Turk (blue) latent score distributions are distributed in similar patterns. The density plot for the Mechanical Turk distribution is smoother because of the much smaller sample size. The Prejudice Index distribution is shifted in the Mechanical Turk

sample because of the absence of any observations of certain scale values in the smaller sample (i.e., strong pro-Black attitudes), which changed the numeric center of the scale.

## General Discussion

Historically, a major pillar of social psychology is theoretical models postulating the structure and consequences of people's attitudes, motivations, or beliefs toward those in other groups. Thousands of papers have been devoted to this topic and numerous scales have been developed to tap these constructs. Accurate measurement is at the core of this theoretical progress. Those studying racial attitudes, motivations, and beliefs need to measure racial attitudes, motivations, and beliefs with precision. To the extent measurement is poor, data cannot provide clear evidence for theoretical models, even in the presence of significant findings. Improved measurement is ultimately critical for knowledge accumulation and without it the field is hindered (Flake & Fried, 2020).

To aid in this process, we performed the most comprehensive evaluation of race-related scales to date, evaluating the validity properties of 25 race-related scales using modern techniques. Modern fit indices (McNeish & Wolf, 2021) found that model fit of most scales ranged from unacceptable to highly unacceptable. Further, we also revealed that some of the best-fitting race-related scales, such as Bayesian Racism and Modern Racism, exhibited problematic "peaks" at the floor of their distributions, indicating these scales are less adept at differentiating between individuals at lower values of the latent construct. Finally, the nomological net helped to identify that measuring prejudice is a saturated space, and aided in our subsequent recommendations for scale use below.

## Recommendations

Simultaneously considering the results of our wide-ranging analyses, four concrete recommendations emerge, along with three additional observations.

### *Recommendation 1*

For researchers who do not have a strong *a priori* reason to use a specific measure of prejudice, we recommend using Prejudice Index, Modern Racism, or Bayesian Racism scale to measure general anti-Black prejudice. This recommendation is grounded in both the superior model fit indices of these constructs in the CFA section as well as their locations in the nomological net. Poor model fit indicates a mismatch between the theoretical structure of the model and the observed data, which makes unclear whether a specific latent construct is being measured at all (as shown in Figure 1). If the data do not fit the theorized structure of data, researchers are not measuring what they think they are measuring, and their conclusions are more likely to be wrong. Modern Racism, Bayesian Racism, and Prejudice Index are located in the central cluster of the nomological net, a cluster that we interpret as "general anti-Black prejudice". Thus, when researchers do not have interest in a specific race-related theoretical construct, we recommend these three scales.

### *Recommendation 2*

We recommend using the Prejudice Index over Modern Racism and Bayesian Racism when researchers wish to differentiate between levels of pro-Black/anti-White sentiment. This recommendation is grounded in the IRT results for both the distribution of latent factor scores and the local reliabilities. Modern Racism and Bayesian Racism both demonstrate a floor effect, poorly capturing variation at the bottom of the scale, whereas the Prejudice Index, which uses difference scores between ratings of Black and White groups, has no such limitation. If a scale's

latent factor score does not cover a particular range of values, the scale is not sensitive to variation of the construct in that area. Conclusions hinging on sensitive measurement in that range are more likely to be wrong. Thus, the Prejudice Index is better suited for answering questions that pertain to variation in pro-Black/anti-White sentiments.

*Recommendation 3*

As a more general recommendation, we reiterate that simply reporting Cronbach's α and McDonald's ω as evidence of a scale's validity is insufficient (see Flake et al., 2017). Many scales with high α and ω scores performed quite poorly in terms of model fit (e.g., Right Wing Authoritarianism and General Intergroup Anxiety). Conversely, Prejudice Index, one of the best-performing scales in terms of model fit and latent score distribution coverage, had α and ω scores that were acceptable but relatively low compared to most other scales. We echo many others in cautioning against authors' use of these scores as standalone justification for an existing or novel scale and correspondingly recommend that editors and reviewers push back against this practice.

*Recommendation 4*

Finally, we emphasize that researchers with strong motivation to measure a specific latent construct should not necessarily hesitate to use the appropriate scale, keeping in mind the potential limitations that come with this decision (e.g., low confidence that the latent construct of interest is actually being captured). In this case, we recommend incorporating scale evaluation as part of the project and considering scale renovation (discussed below).

*Additional Observations*

Researchers seeking to measure motivations to control prejudice would be reasonably well-served by the Internal and External Motivation to Control Prejudice scales, with a couple caveats. The scale overall shows decent but not good model fit, and although External

Motivation appears to be a quite distinct latent construct, Internal Motivation appears to be part of the general anti-Black prejudice cluster of scales in the nomological net.

Researchers seeking to measure cultural knowledge might be better served by the items in the Cultural Attitudes Toward Black People or Perceptions of Others' Prejudice scales if they regard them as separate indicators of cultural knowledge about specific traits (e.g., aggression, attractiveness, trustworthiness). These scales do not appear to capture a single underlying latent construct.

Finally, the high correlations between many of the scales in the nomological net suggests that, in general, the theoretical space related to racial stereotyping and prejudice is highly saturated. We recommend that researchers think carefully about the extent to which a given scale that purports to measure a specific kind of racial prejudice or race-related attitude actually does so, at least in a way that is theoretically distinct from other related attitudes. If it is the case that some of these scales are conceptually redundant, this justifies the selection of scales for their measurement properties.

**What This Work Does Not Mean**

Although we present concrete recommendations, we also wish to be clear about what we are *not* saying. First, all of the recommendations above are based solely on the scales' psychometric properties and location in the nomological network. External validity evidence for these constructs was not the aim of the present research, and we cannot speak to how well these scales predict outcomes of interest (though, all else equal, scales with more measurement error are less likely to predict with precision). Some researchers might believe that a specific scale is particularly well-suited for predicting a certain outcome. Although a scale's central position in the nomological net might cast some doubt on the potentially unique ability of a specific scale to

predict a certain outcome, scales centrally located in the nomological net nevertheless possess some variance that is unique from other scales. In these cases, researchers can look to theory and previous external validity evidence for guidance.

Second, we are agnostic to the historical structural validity of these scales. Many of the scales evaluated are more than 20 years old and may have shown different psychometric properties when initially developed. In fact, part of our justification for the present research is that construct validation is an ongoing and living process (Cronbach & Meehl, 1955), such that both the content validity of individual items and the research culture broadly shift over time. Researchers will continue to ask different kinds of questions about different populations in different contexts. Some of these scales were originally administered to samples of college students, who are a considerably more constrained population than that sampled here. Finally, actual racial attitudes and beliefs also shift over time (Charlesworth & Banaji, 2019; Devine & Elliot, 1995), which may explain why we observed floor effects for many of the scales.

Third, we are not claiming that any scales reviewed here are uninformative. Although we do find that many of the scales are "noisy" instruments for measuring latent factors, some signal is captured. Researchers have revealed myriad important findings regarding stereotyping, prejudice, and discrimination using many of the scales reviewed in this paper, and we certainly do not argue that these findings are invalid. Rather, we view these results through an optimistic lens, as a guide for both selecting current best scales and for identifying useful avenues for scale renovation. To this end, we hope that our analyses lead to future work seeking to create updated versions of these scales that address some of the measurement weaknesses identified here.

Finally, we want to note that issues with the structural validity of psychological scales are not unique or specific to race-related scales. Although we focus on evaluating these scales, it is

likely the case that many scales across the social and personality literature exhibit similar issues (e.g., Hussey & Hughes, 2020).

**Implications for Scale Development and Renovation**

This work highlights some clear future directions for scale development in racial stereotyping and prejudice research. Some areas of the nomological network are relatively sparse and feature few or no scales that show good structural validity. Researchers interested in investigating effects of stereotype knowledge or motivation to control prejudice might see this as an opportunity to develop a new scale using modern methods, which would constitute a valuable improvement.

Furthermore, the information provided by IRT about individual items (available on the OSF page) is an excellent resource for systematically renovating existing scales, providing two main benefits for scale renovation. First, IRT analyses identify weak items that provide limited information to the latent factor (similar to examining latent factor loadings in CFA). For example, the IRT results for Right Wing Authoritarianism show that there are two items in particular that provide low information about the latent factor and could be removed with little loss. Second, IRT analyses identify the range of the latent factor at which each item is informative, allowing researchers to identify when introducing a "harder" item (i.e., one that discriminates between those very high in the latent factor) or an "easier" item (i.e., one that discriminates between those very low in the latent factor) would improve the coverage of a scale. For example, although the Modern Racism scale shows very good model fit, IRT results suggest that the addition of a few more extreme pro-Black items would improve the coverage of the scale and differentiate between the high percentage of individuals who hit the floor of the scale. A figure illustrating these examples is available on the OSF page.

We certainly do not suggest abandoning rich theoretical constructs such as Right Wing Authoritarianism or Symbolic Racism; rather, we suggest that there is great opportunity for renovating these scales, which will improve future research on these topics. We suggest that researchers interested in scale renovation employ IRT to identify uninformative items to cut and the difficulty level at which new items should be introduced. Overall, we hope that this work motivates and rewards researchers who pursue scale renovation and believe that such work would be highly beneficial to the field and to the further development of theories that hinge on the accurate measurement of specific latent constructs.

**Limitations**

We note a few key limitations of the current work. First, some of the scales used in our sample already have recently renovated versions that were not collected in our analysis. We note two prominent cases here. First, the SDO7 (Ho et al., 2015) renovates the scale items and reconceptualizes Social Dominance Orientation as a two-dimensional construct. SDO6 was used in the present work (Pratto et al., 1994), but it is important to note that this scale is still regularly used. For example, between January and March 2021, we identified seven papers published using SDO6. Similarly, the Racial Resentment scale was renovated in 2011 (Wilson & Davis, 2011), and our analyses reflect the psychometric properties of an earlier version (Kinder et al., 1996). Future analyses might include these updated versions of the scales, but our findings here are relevant to modern research even for these older but still used scales.

Our evaluation of race-related scales also does not capture the full “universe” of scales available in the literature. One notable exclusion (due to its absence from the Project Implicit dataset) is the Color-Blind Racial Attitudes Scale (Neville et al., 2000), which has been cited

over 1100 times. Future work might collect or use data that includes important scales absent from the current investigation.

Furthermore, we used a self-selected sample of individuals who chose to visit Project Implicit and Mechanical Turk. It is possible that these scales show different psychometric properties in different populations. However, we note our analyses are already on a far larger and more diverse population than the original scale development work, which used smaller and more homogenous samples of American adults, White adults, and college students.

We have evaluated the construct validity of these scales with regard to measuring racial attitudes toward Black people among mostly U.S. participants. Because construct validation pertains to a specific *use* of a scale and can be context or population dependent (Kane, 2013; Messick, 1995), it is not necessarily the case that scales with good psychometric properties in this scenario would have good properties when assessing attitudes toward other groups drawing from other populations. Researchers using these scales should always first verify their measures have properties similar to previous analyses to ensure their measures are working as expected, especially for any new context.

**The Ongoing Theoretical and Methodological Importance of Explicit Bias**

Finally, we suggest that it may be a suitable time to revitalize research on explicitly expressed prejudice. Beginning in the 1980s, social scientists were increasingly concerned that individuals were no longer honestly reporting their prejudices on explicit self-report measures, due to social desirability concerns and the idea that appearing prejudiced was no longer publicly acceptable. Accordingly, the field began developing indirect assessments of bias (Devine, 1989; Fazio et al., 1986; Gaertner & McLaughlin, 1983; Greenwald et al., 1998). This focus fueled nearly 40 years of intense research into indirectly measured implicit biases, what they are, their

causes, their consequences (Cameron et al., 2012; Dovidio et al., 2002; Greenwald et al., 2009; Hofmann et al., 2005; Kawakami et al., 2007; Kurdi et al., 2018; Nosek et al., 2007; Payne et al., 2005). This research has greatly informed our understanding of social cognition and bias, yet has also revealed some of the limitations of indirectly measured biases. Like all cognitive tasks, they have high measurement error (Cunningham et al., 2001; Gawronski et al., 2017; Hedge et al., 2018) and only weak relationships with behavior (Greenwald et al., 2009; Kurdi et al., 2018; Oswald et al., 2013). In contrast, explicit measures of racial attitudes have less measurement error (Gawronski et al., 2017) and stronger or at least equivalent relationships with individual level behavior (Oswald et al., 2013). Although we understand concerns about socially desirable responding, we do not believe there is a shortage in modern times of public expressions of prejudice (Crandall et al., 2018).

In all, explicit measures have superior measurement properties relative to implicit measures of racial attitudes. They have, at best, equal associations with behavior, yet explicit biases are easier to measure. People also appear to be willing to explicitly express prejudice toward stigmatized groups. Accordingly, we believe the need for effective self-report measures of explicit bias is alive and well, and encourage prejudice researchers to continue empirical attention on explicitly endorsed measures of racial prejudice and collect alongside implicit measures. The analyses provided in the present research can aid this endeavor.

**Conclusion**

Before any deep-sea dive, researchers and engineers carefully test their equipment to make sure that every tool and instrument is functioning properly. Although psychologists are not faced with the same high-cost, life-threatening stakes, we can nevertheless benefit by following suit, carefully considering and testing the instruments we use to study racial attitudes and other

latent factors. By closely evaluating the measurement scales we use to “dive” into the minds of others and reveal people’s thoughts and beliefs, we can come ever closer to actually observing these thoughts and beliefs, allowing us to draw stronger conclusions about their nature, meaning, and consequences.

## Open Practices Statement

The data and materials for all studies are available at https://osf.io/zg6fr/?view_only=e6d56172a5934259a81729312ebf0754.

## References

Axt, J. R. (2018). The best way to measure explicit racial attitudes is to ask about them. *Social Psychological and Personality Science, 9*(8), 896–906. https://doi.org/10.1177/1948550617728995

Baker, F. B. (2001). *The basics of item response theory*. ERIC Publications.

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences, 42*(5), 815–824. https://doi.org/10.1016/j.paid.2006.09.018

Cameron, C. D., Brown-Iannuzzi, J. L., & Payne, B. K. (2012). Sequential priming measures of implicit social cognition: A meta-analysis of associations with behavior and explicit attitudes. *Personality and Social Psychology Review, 16*(4), 330–350. https://doi.org/10.1177/1088868312440047

Charlesworth, T. E. S., & Banaji, M. R. (2019). Patterns of implicit and explicit attitudes: I. Long-term change and stability from 2007 to 2016. *Psychological Science, 30*(2), 174–192. https://doi.org/10.1177/0956797618813087

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology, 83*(6), 1314–1329. https://doi.org/10.1037//0022-3514.83.6.1314

Correll, J., Wittenbrink, B., Crawford, M. T., & Sadler, M. S. (2015). Stereotypic vision: How stereotypes disambiguate visual stimuli. *Journal of Personality and Social Psychology, 108*(2), 219–233. https://doi.org/10.1037/pspa0000015

Crandall, C. S., Miller, J. M., & White, M. H. (2018). Changing norms following the 2016 U.S. presidential election: The Trump effect on prejudice. *Social Psychological and Personality Science, 9*(2), 186–192. https://doi.org/10.1177/1948550617750735

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *Interjournal, Complex Systems*, *1695*(5), 1–9.

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological Science*, *12*(2), 163–170. https://doi.org/10.1111/1467-9280.00328

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*(1), 5–18.

Devine, P. G., & Elliot, A. J. (1995). Are racial stereotypes really fading? The Princeton trilogy revisited. *Personality and Social Psychology Bulletin*, *21*, 1139–1150.

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology*, *82*(1), 62–68. https://doi.org/10.1037//0022-3514.82.1.62

Dudek, F. J. (1979). The continuing misinterpretation of the standard error of measurement. *Psychological Bulletin*, *86*(2), 335–337. https://doi.org/10.1037/0033-2909.86.2.335

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*(3), 399–412. https://doi.org/10.1111/bjop.12046

Duriez, B., & Soenens, B. (2009). The intergenerational transmission of racism: The role of Right-Wing Authoritarianism and Social Dominance Orientation. *Journal of Research in Personality*, *43*(5), 906–909. https://doi.org/10.1016/j.jrp.2009.05.014

Edwards, M. C. (2009). An introduction to Item Response Theory using the Need for Cognition scale. *Social and Personality Psychology Compass*, *3*(4), 507–529. https://doi.org/10.1111/j.1751-9004.2009.00194.x

Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science*, *16*(4), 779–788. https://doi.org/10.1177/1745691620970586

Fabrigar, L. R., & Wegener, D. T. (2016). Conceptualizing and evaluating the replication of research results. *Journal of Experimental Social Psychology*, *66*, 68–80. https://doi.org/10.1016/j.jesp.2015.07.009

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, *50*(2), 229–238. https://doi.org/10.1037/0022-3514.50.2.229

Feldman, S., & Huddy, L. (2005). Racial resentment and White opposition to race-conscious programs: Principles or prejudice? *American Journal of Political Science*, *49*(1), 168–183. https://doi.org/10.1111/j.0092-5853.2005.00117.x

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Flora, D. B. (2020). Your coefficient alpha is probably wrong, but which coefficient Omega is right? A tutorial on using R to obtain better reliability estimates. *Advances in Methods*

*and Practices in Psychological Science*, 2515245920951747. https://doi.org/10.1177/2515245920951747

Freeman, J. B., Stolier, R. M., & Brooks, J. A. (2020). Dynamic interactive theory as a domain-general account of social perception. In *Advances in Experimental Social Psychology* (Vol. 61, pp. 237–287). Elsevier. https://doi.org/10.1016/bs.aesp.2019.09.005

Gaertner, S. L., & McLaughlin, J. P. (1983). Racial stereotypes: Associations and ascriptions of positive and negative characteristics. *Social Psychology Quarterly*, *46*(1), 23–30.

Gawronski, B., Morrison, M., Phills, C. E., & Galdi, S. (2017). Temporal stability of implicit and explicit measures: A longitudinal analysis. *Personality and Social Psychology Bulletin*, *43*(3), 300–312. https://doi.org/10.1177/0146167216684131

Ghavami, N., & Peplau, L. A. (2013). An intersectional analysis of gender and ethnic stereotypes: Testing three hypotheses. *Psychology of Women Quarterly*, *37*(1), 113–127. https://doi.org/10.1177/0361684312464203

Greenwald, A. G., Mcghee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*, 1464–1480.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*(1), 17.

Hancock, G. R., & Mueller, R. O. (2011). The reliability paradox in assessing structural relations within covariance structure models. *Educational and Psychological Measurement*, *71*(2), 306–324. https://doi.org/10.1177/0013164410384856

Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But…. *Communication Methods and Measures*, *14*(1), 1-24.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

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

Hester, N., & Gray, K. (2018). For Black men, being tall increases threat stereotyping and police stops. *Proceedings of the National Academy of Sciences*, 201714454. https://doi.org/10.1073/pnas.1714454115

Hiel, A. V., & Mervielde, I. (2005). Authoritarianism and Social Dominance Orientation: Relationships with various forms of racism. *Journal of Applied Social Psychology*, *35*(11), 2323–2344. https://doi.org/10.1111/j.1559-1816.2005.tb02105.x

Ho, A. K., Sidanius, J., Kteily, N., Sheehy-Skeffington, J., Pratto, F., Henkel, K. E., Foels, R., & Stewart, A. L. (2015). *The nature of Social Dominance Orientation: Theorizing and measuring preferences for intergroup inequality using the new SDO7 scale*. *109*(6), 1003–1028.

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. *Personality and Social Psychology Bulletin*, *31*(10), 1369–1385. https://doi.org/10.1177/0146167205275613

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424–453. https://doi.org/10.1037/1082-989X.3.4.424

Hussey, I., & Hughes, S. (2020). Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science*, *3*(2), 166–184. https://doi.org/10.1177/2515245919882903

Kamada, T., & Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters*, *31*(1), 7–15.

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

Kawakami, K., Phills, C. E., Steele, J. R., & Dovidio, J. F. (2007). (Close) distance makes the heart grow fonder: Improving implicit racial attitudes and interracial interactions through approach behaviors. *Journal of Personality and Social Psychology*, *92*(6), 957–971. https://doi.org/10.1037/0022-3514.92.6.957

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A. G., & Banaji, M. R. (2018). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *American Psychologist*, *74*(5), 569. https://doi.org/10.1037/amp0000364

Li, C.-H. (2016a). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936–949. https://doi.org/10.3758/s13428-015-0619-7

Li, C.-H. (2016b). The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. *Psychological Methods, 21*(3), 369. https://doi.org/10.1037/met0000093

McDonald, R. P. (2013). *Test theory: A unified treatment*. psychology press.

McNeish, D., An, J., & Hancock, G. R. (2018). The thorny relation between measurement quality and fit index cutoffs in latent variable models. *Journal of Personality Assessment, 100*(1), 43–52. https://doi.org/10.1080/00223891.2017.1281286

McNeish, D., & Wolf, M. (2020). *Dynamic Model Fit. R Shiny application version 1.1.0*.

McNeish, D., & Wolf, M. G. (2021). Dynamic fit index cutoffs for confirmatory factor analysis models. In *Psychological Methods*. https://doi.org/10.31234/osf.io/v8yru

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Millsap, R. E. (2007). Structural equation modeling made difficult. *Personality and Individual Differences, 42*(5), 875–881. https://doi.org/10.1016/j.paid.2006.09.021

Nicol, A. A. M., & Rounding, K. (2013). Alienation and empathy as mediators of the relation between Social Dominance Orientation, Right-Wing Authoritarianism and expressions of racism and sexism. *Personality and Individual Differences, 55*(3), 294–299. https://doi.org/10.1016/j.paid.2013.03.009

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., Smith, C. T., Olson, K. R., Chugh, D., Greenwald, A. G., & Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology, 18*(1), 36–88. https://doi.org/10.1080/10463280701489053

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology, 105*(2), 171–192. https://doi.org/10.1037/a0032734

Pauker, K., Meyers, C., Sanchez, D. T., Gaither, S. E., & Young, D. M. (2018). A review of multiracial malleability: Identity, categorization, and shifting racial attitudes. *Social and Personality Psychology Compass, 12*(6), e12392. https://doi.org/10.1111/spc3.12392

Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology*, *81*(2), 181.

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, *89*(3), 277–293. https://doi.org/10.1037/0022-3514.89.3.277

Rizopoulos, D. (2006). **ltm**: An *R* Package for Latent Variable Modeling and Item Response Theory Analyses. *Journal of Statistical Software, 17*(5). https://doi.org/10.18637/jss.v017.i05

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5—12 (BETA). *Journal of Statistical Software*, *48*(2), 1–36.

Saucier, D. A., & Miller, C. T. (2003). The persuasiveness of racial arguments as a subtle measure of racism. *Personality and Social Psychology Bulletin*, *29*(10), 1303–1315. https://doi.org/10.1177/0146167203254612

Sears, D. O. (1988). Symbolic racism. In *Eliminating racism* (pp. 53-84). Springer, Boston, MA.

Tighe, J., McManus, I., Dewhurst, N. G., Chis, L., & Mucklow, J. (2010). The standard error of measurement is a more appropriate measure of quality for postgraduate medical

assessments than is reliability: An analysis of MRCP(UK) examinations. *BMC Medical Education*, *10*(1), 40. https://doi.org/10.1186/1472-6920-10-40

Tosh, C., Greengard, P., Goodrich, B., Gelman, A., Vehtari, A., & Hsu, D. (n.d.). *The piranha problem: Large effects swimming in a small pond*. 16.

Wilson, D. C., & Davis, D. W. (2011). Reexamining racial resentment: Conceptualization and content. *The ANNALS of the American Academy of Political and Social Science*, *634*(1), 117–133. https://doi.org/10.1177/0002716210388477

Zanon, C., Bastianello, M. R., Pacico, J. C., & Hutz, C. S. (2013). Development and validation of a positive and negative affect scale. *Psico-USF*, *18*(2), 193–201. https://doi.org/10.1590/S1413-82712013000200003

Zanon, C., Hutz, C. S., Yoo, H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexão e Crítica*, *29*(1), 18. https://doi.org/10.1186/s41155-016-0040-x