Journal of Methods and Measurement in the Social Sciences,
Vol. 12, No. 2, 95-116, 2021

The Modern Biased Information Test: Proposing alternatives for implicit measures

Aurelio José Figueredo
University of Arizona

Vanessa Smith-Castro
Universidad de Costa Rica

Mateo Peñaherrera-Aguirre
University of Arizona

The present article describes the development of a Modern Biased Information Test (MBIT) inspired by the work published by Donald Campbell in 1950 on indirect measures of prejudice. A biased information test aims to tap individuals' intergroup attitudes from the selective information they use to describe group members. Two biased information tests were developed to measure ethnocentric and androcentric biases, respectively, and applied in four convenience samples of students from two different cultural settings (Costa Rica and the USA). The internal consistency for the *accuracy* indicators derived from both tests was acceptable and comparable across cultures. In contrast, the internal consistency for ethnocentric *biases* was adequate across samples and cultures, but the internal consistency for androcentric *biases* was unacceptable across both cultures. Results are discussed in the line of the usefulness of alternative measures for tapping implicit attitudes.

Keywords: Implicit Bias, Geometric Mean, Generalizability Theory Analysis, Ethnocentric Bias, Androcentric Bias

The use of implicit measures in social psychology has increased exponentially in the last decades (Gawronski, 2019). Researchers in the intergroup relations field, for instance, enthusiastically received these new tools as a promising way to capture racial biases that have indeed been reduced in the last years but also have become less blatant, more subtle, highly ambivalent, and hardily sanctioned by social norms. Yet, these measurement strategies have also been subject to fierce criticism in the light of the fact that their test-retest stability is low, their predictive value for behavior is weak, and their relationship with explicit measures is unstable and not clearly explained (Brownstein et al., 2020; Greenwald et al., 2020).

The current debate around the shortcomings of implicit measures had led some scholars to raise severe doubts about their usefulness (Blanton et al., 2009; Fiedler et al., 2006; Oswald et al., 2013). However, for other authors, these unsatisfactory results are not sufficient to cast the whole implicit bias research into doubt and recommend the continued evaluation of the current measures as well as the development of different ways to capture race bias in an indirect way (Gawronski, 2019; Meissner et al., 2019).

In line with this spirit, the present article presents an alternative way to measure implicit bias following Campbell's (1950) suggestion of developing tests that measure bias from the information that people claim to have about outgroups. This research illustrates the development of these types of tests and provides the first psychometric information of the instruments when applied to tap ethnocentric and androcentric biases in two different cultural settings. The basic principle underlying Campbell's (1950) original *Biased Information Test (BIT)* is that the construct of *prejudice* against any particular group would be reflected in biased *beliefs* regarding the characteristics of the biocultural population in question. The BIT was therefore a test of value-laden information regarding the characteristics of the other group. This information was thus not just neutral facts, but facts with judgmental value, comprising characteristics generally held to be either socially desirable or undesirable. The attribution of negatively-valued characteristics would therefore indicate a systematic bias against the target group; the attribution of positively-valued characteristics would therefore indicate a systematic bias favoring the target group. Although emotions themselves are not directly assessed, it is therefore the value-laded content of the information tested that is designed to reveal either positive or negative prejudice.

The *Modern Biased Information Test (MBIT)* being introduced and evaluated in the present article is based directly upon the *Biased Information Test (MBIT)* introduced by Donald Campbell in 1950. The most fundamental reason for creating the *Modern* version is that both the original BIT and the derived MBIT are based on biodemographic population parameters which change over time, in particular over the past 70 years. In addition, some of the terms used in 1950 to describe certain ethnic groups (e.g., "Negroes") are generally considered offensive today, although they were not at the time (for example, it was a term used without any opprobrium by the Reverend Martin Luther King himself). It is thus necessary to tailor any version of the MBIT designed to measure prejudice for or against any contemporary ethnic group to the correct and updated biodemographic population parameters for as close to the time of test administration as possible.

Measuring intergroup bias

The measurement of intergroup phenomena as prejudice, stereotypes, and discrimination has been quite challenging since the beginning of this line of research. The first papers addressing the measurement of prejudice were published in the early 1920s. In 1924 Robert Park proposed social distance to operationalize prejudice. As he claimed: "What we ordinarily call prejudice seems then to be more or less instinctive and spontaneous disposition to maintain social distances" (p. 343). A year later, Emory Bogardus (1925) presented a procedure to measure social distance employing a variation of what was later known as Guttman's (1944) rank order scaling system and developed several intergroup contact and distance indexes. According to Bogardus, these indexes "... do not indicate merit or traits of the respective races, but rather something of the extent of the social contacts open to each race... The social contact range all indicates something regarding the racial attitudes of the raters" (p. 302).

Bogardus' Social Distance Scale exemplifies very well what an explicit measure for assessing racial bias is. In this procedure, participants are asked to self-disclose their feelings of acceptance for outgroup members. Specifically, the scale asks respondents to report whether they would relate to members of an outgroup in various ways, ranging from "accepting them as close relatives by marriage" to "excluding them from my country." In this way, the scale measures respondents' perceived sense of intimacy or closeness to those groups that are different from their own. As in the Bogardus Social Distance Scale, items of a typical prejudice scale are verbal statements or phrases that participants should evaluate based on their previous stored attitudes and knowledge. Once participants retrieved the necessary information, they must decide how to better express their response based on the offered numerical scale. The cognitive processes of answering prejudice items are similar to those demanded to answer Semantic Differentials or Feelings Thermometers, i.e. the respondent is supposed to understand the task, recall the relevant information to answer it, make a judgment, and find the *proper* respond (Turengau, 1984).

Thus, explicit (direct, self-report) measures rely on the willingness and capacity of individuals to self-report conscious and controlled cognitive processes, and these features have been associated with several limitations as self-presentation bias, social desirability effects, and simulation of test results (Olson & Zabell, 2016). For instance, data show that participants in a more relaxed context express less racial bias against Arabs than participants in a “cognitive-load and time-pressure” condition (Echabarria Echabe, 2013), suggesting that the overt expression of racial bias occurs under conditions in which less control can be exerted.

Given the well-known limitations of traditional explicit measures (i.e., self-presentation, social desirability, and simulation of test results), implicit measures have been developed as complements and, in some cases, as substitutes of traditional scales (Fazio & Olson, 2003). Implicit tests include a wide range of techniques as sitting distance (Mann & Kawakami, 2012), arguments judgment tasks (Heitland & Bohner, 2010), word or sentence fragments completion tasks (Koopman et al., 2013), advice-taking tasks (Webb, 2011), affective priming techniques (Fazio et al. 1986), response times (Greenwald et al., 1998), and physiological activity as blink startle, heart rate, vascular performance, or brain activity among others (i.e., Amodio et al., 2003, Harris & Fisle, 2007). Thus, implicit measures intend to tap individuals' racial attitudes when they do not have the goal to express them (i.e., unintentional) or when they have the goal to conceal them (i.e., uncontrollable) (Fazio & Olson, 2003).

The Implicit Association Test (IAT) (Greenwald et al., 1998; 2002) is perhaps the most well-known example of an implicit measure. In this task, participants first perform two categorization tasks (e.g., labeling names as Black or White, classifying words as positive or negative) separately, and then the tasks are combined. Based on the assumption that people respond faster to more associated pairs of stimuli, the speed of decisions to specific combinations of words (e.g., Black names with unpleasant attributes vs. Black names with pleasant characteristics) represents the implicit measure. The logic of the procedure is that when the pairing is congruent in the participant's mind (e.g., Black and unpleasant), the reaction time of making the classification is shorter.

As any attempt to measures psychological constructs, implicit measures are not exempt from limitations. Several longitudinal studies have shown that individuals' scores on implicit measures, especially when measuring racial attitudes, are unstable (Gawronski et al., 2017). Recently, Greenwald and collaborators (2020) reported test-retest correlations of IAT scores on sensitive topics as stereotypes around r = .50 among adults, while Rae & Olson (2018) reported mean test-retest reliability *rs* between .34 and .48 among children. Studies also show that the average correlations between implicit measures and behavior are weak. For example, Oswald and collaborators' (2013) meta-analysis of 46 studies between 2004 and 2011 estimated the average correlation between IAT scores and racial and ethnic discrimination measures between r = .12 and r = .15. These results indicate that individual differences in implicit bias account for approximately 2% of the variance in intergroup discrimination. Finally, studies also show a low correlation between implicit and explicit measures (Echaberría Echabe, 2013). Hofmann et al. (2005) investigated the correlation between IAT and explicit self-report measures in a meta-analytic study on a sample of 126 independent studies. The authors found a mean effect size of r = .24, with approximately half of the variability of the correlations attributable to moderator variables.

These unsatisfactory results have been the subject of an intense debate. For some authors, these inconsistencies reflect the many available methods designed to address different aspects of the responding processes (e.g., awareness or automaticity) and different underlying systems or mechanisms (e.g., conceptual or perceptual; Goodall, 2011). Other authors call attention to the fact that some researchers use the term implicit in describing the measurement procedures, whereas others use it to describe the nature of the assessed psychological construct (see Fazio and Olson, 2003 for the analysis of the distinction).

The Associative-Propositional Evaluation (APE) model developed by Gawronski and Bodenhausen (2006, 2007) is a good example of this distinction. These authors assume the existence of two independent constructs, namely implicit and explicit attitudes. The former are affective automatic reactions aroused by encounters with an object. According to them, implicit attitudes are shaped and changed via associative processes due to the pairing with another positive or negative stimulus. In contrast, explicit attitudes are conscious evaluations of an object formed and changed by the availability of information about that object. Therefore, implicit association or priming tests are more suitable for tapping implicit attitudes, while direct questions or statements are more suitable to measure explicit attitudes. Other dual-attitude models (e.g., Devine, 1989) also propose different constructs underlying explicit and implicit measures. According to these models, people hold multiple attitudes towards a topic simultaneously with the newer attitudes layered on top of the older ones. When people retrieve their attitudes, they explicitly report the most current ones, but the older ones can be measured only with implicit techniques (Echaberría Echabe, 2013).

Finally, and this is perhaps the most important critique, some scholars point out the lack of solid theoretical models explaining what causes implicit scores making it impossible to know what constructs are actually measured. Fiedler et al. (2006) argue, for example, that it is not possible to infer from an observed IAT effect that there is an underlying negative attitude that explains the effect, simply because many other factors, besides genuine attitudes, can produce IAT effects.

Indeed, some research on young adult samples has shown that IAT effects may be better understood as produced by intergroup phenomena other than prejudice. For instance, van Ravenzwaaijwe and collaborators (2011) conducted three different IATs with Dutch undergraduates in which the race and the names of the stimuli were varied. In one such study, the use same-race Dutch names versus racially charged Moroccan names; in a second study, they employed same-race Dutch names versus racially neutral Finnish names; and in the third study they used Moroccan names versus Finnish names. van Ravenzwaaijwe and collaborators found that participants responded similarly to the racially charged outgroup Moroccan names and the racially neutral outgroup Finnish names, but when ingroup Dutch names were contrasted with either of the two outgroup names, there was an IAT effect. Thus, the results of these experiments offer no support for the contention that the name-race IAT originates mainly from a prejudice based on race, they rather support the alternative explanation that the IAT effect is due to ingroup-outgroup membership.

In this context, the revision of the current methods and the continuous effort to design new strategies of assessment seems urgent for many reasons. First, research often requires the repeated assessment of prejudice over multiple time points, especially when interventions or changes in policy need to be evaluated, making necessary the development of new instruments for measure bias. Second, studies have shown that the expression of prejudice has reduced over time and has become more subtle and ambivalent. These results suggest finding better ways to track these changes over time and measure their effects. Third, many implicit measures are based on response latencies requiring special hardware, equipment, and adequate conditions in laboratory settings. Therefore, less expensive measures of racial bias (i.e., paper-and-pencil questionnaires) would likely be a valuable addition.

The present article describes the development of a Modern Biased Information Test (MBIT) inspired by the work published by Donald Campbell in 1950 on indirect measures of prejudice. As its name suggests, a biased information test aims to tap individuals' intergroup attitudes from the selective information they use to describe group members. Following this premise, we developed two biased information tests to measure ethnocentric and androcentric biases, respectively. We then applied the tests in four convenience samples of students from two different cultural settings (Costa Rica and the USA). In this report, we describe the development of these measures and provide the first evidence of the internal consistency of their scores.

Biased information tests

In his work "The indirect assessment of social attitudes," Campbell (1950) offered an examination of several techniques for measuring attitudes popular in the academic community of that time, including projective tests, doll play techniques, sentence completion tasks, and information tests. The latter was diagnostic for intergroup attitudes derived from people's systematic information biases about members of specific social or ethnic groups. Campbell (1950) described this type of test as “disguised-structured tests of social attitudes”.

The Information Tests reviewed by Campbell asked participants to evaluate items with information of members of different social groups. Examples of these items are: "Jews form about 25% of the Communist Party", "There are no Negro Congressmen today," "[The] average weekly wage of the war worker in 1945 was..." (see Hamond, 1948 and Loeblowitz-Lennard & Riessman, 1946, for a detailed description of the tests). Some tests included true-false items; others offered one response option with the factual information and one or more false options; in other cases, all options were incorrect but in opposite directions.

Although different in format, all tests shared the assumption that the propensity of individuals to selectively recall, activate and look for information that confirms their believes, stereotypes, and opinions will reveal their attitudes in favor or against the target groups. As Campbell (1950) pointed out: "You would all probably agree that in a detailed test of information, the direction of people's guesses or misconceptions will frequently bear a relationship to their attitudes. In a complementary fashion, a given person's knowledge is apt to reflect in its unevenness his selective awareness and retention, or his biased sources of information" (p. 20). This phenomenon was defined later as confirmation bias by P.S. Wason (1960).

As pointed out before, we constructed two MBIT, one aimed to measure bias towards immigrants (MBIT-ethnocentrism), and the other intended to measure bias against women (MBIT-androcentrism). Once assembled, questionnaires were reviewed via Cognitive Interview (Willis, 2005, Smith-Castro & Molina-Gallardo, 2011) to detect and correct comprehension problems. The final versions were applied in four different Costa Rica and the USA samples, as described in the following sections.

Method

Study participants

In Costa Rica, a total of 131 university students, 52% females (age *M* = 22 years; *SD* = 5.62), answered the MBIT-ethnocentrism items, and 150 university students, 50% females (age *M* = 23 years; *SD* = 4.68), answered the MBIT-androcentrism items. In USA, a total of 134 Southwestern university students, 70% females (age *M* = 19 years; *SD* = 1.94) answered the MBIT-ethnocentrism items, and 199 Southwestern university students, 70% females (age *M* = 20 years; *SD* = 4.39), answered the MBIT-androcentrism items.

Measures and Procedures

The MBIT-ethnocentrism instructed participants as follows: "These next questions aim at finding out how adequate your knowledge about Nicaraguan / Mexican immigrants is compared to Costa Rican/USA citizens. Data come mostly from the Census, household surveys, and official statistics of Costa Rican/USA public institutions of the last five years. Even if you don't know the answers with absolute certainty, we ask you to make your best guess for each question. Please circle the letter corresponding to your response".

The MBIT-androcentrism used an analogous instruction to evaluate the "knowledge" of the participants about men and women. Both measures covered five domains: Health, education, work, general demographics, and crime. Examples of the Costa Rican MBIT-ethnocentrism items are: "What percentage of Costa Ricans/Nicaraguan immigrants had not completed Primary school?" or "What percentage of all Costa Rican/Nicaraguan immigrants committed a crime of any kind in 2006?". We ensembled mirror versions of the questionnaires for CR and USA with the same questions. The response options, of course, vary for each country, depending on the actual data gathered from official documents of institutions in each country as the Census Bureau, the National Center of Health Statistics, and the Federal Bureau of Investigation in the USA, or the National Institute for Census and Statistics, the Judicial Investigation Department, and the Ministry of Health in Costa Rica.

Our MBITs differ from the original versions in several ways. First, we used six-option multiple choice answer scales with the correct option in the middle of the scale (options 3 and 4) and distractors equidistant from the correct option in opposite directions. I this way we were able to calculate the magnitude and valence of the deviations from the correct answer and the level of (in)accuracy of the answers.

Second, we generated items with information on the target groups (immigrants/women) and on the reference group (nationals/men). In this way, we can separate bias from general misinformation or guessing tendencies by calculating the difference between the deviations (from the actual answer) of the target group and the deviations from the reference group (the greater the deviations gap, the greater the bias).

Third, we included items offering positive (i.e., rates of higher education attendance among immigrants) and negative information (i.e., rates of incarcerations among women) about the targeted groups. To verify the valence of the information and eliminate ambiguous items, we asked 30 Costa Rican university students in a pilot study to rate their impressions about the people (men and women) who have the traits o is involved in the situations described in each item on a scale from -3 (a very negative impression) to 3 (a very positive impression).

Data scoring

The *raw deviations* indicate the distances, in either the positive or negative direction, between the participants' responses and the "true scores" obtained from our archival sources (e.g., census information) on each of the relevant biodemographic parameters. A positive deviation score indicates that a participant has overestimated the trait; a negative deviation score indicates that a participant has underestimated the trait. If the trait in question is one that is negatively socially valued (e.g., unemployment or poverty), then a positive deviation would indicate that the participant's internal cognitive representation is biased *against* the target group. If the trait in question is one that is positively socially valued (e.g., education or achievement), then a positive deviation would indicate that the participant's internal cognitive representation is biased *in favor of* the target group. Table 1 shows the SPSS syntax for this procedure:

Table 1
SPSS syntax for calculating raw deviations by subtracting the correct response (Option 3 or 4) from the participant's response

SPSS Syntax	Mathematical explanation
COMPUTE MBIT01r = MBIT01-3.	Subtract the correct response (3) from the participant's response on MBIT item number 1.
COMPUTE MBIT03r = MBIT03-3.	Subtract the correct response (3) from the particpant's response on MBIT item number 3.
COMPUTE MBIT06r = MBIT06-4.	Subtract the correct response (4) from the particpant's response on MBIT item number 6.
COMPUTE MBIT07r = MBIT07-3.	Subtract the correct response (3) from the particpant's response on MBIT item number 7.
COMPUTE MBIT09r = MBIT09-4.	Subtract the correct response (4) from the particpant's response on MBIT item number 9.
…	Continue with the rest of the items

The *accuracy estimates* indicate the degree to which these raw deviations were displaced from the "true score" in terms of absolute value, disregarding whether the biases were positive or negative. These scores tell us how much the participants were actually aware of the biodemographic characteristics of the target group regardless of any systematic positive or negative biased that they might have held. Table 2 shows the SPSS syntax for this procedure:

Table 2
SPSS syntax for calculating accuracy estimates as the square root of the squared deviation.

SPSS Syntax	*Mathematical explanation*
COMPUTE MBIT01a = SQRT(MBIT01r*MBIT01r).	Take the square root of the squared deviation of MBIT responses to item 1 to estimate the of absolute value of the deviation score
COMPUTE MBIT03a = SQRT(MBIT03r*MBIT03r).	Take the square root of the squared deviation of MBIT responses to item 3 to estimate the of absolute value of the deviation score
COMPUTE MBIT06a = SQRT(MBIT06r)*(MBIT06 r).	Take the square root of the squared deviation of MBIT responses to item 6 to estimate the of absolute value of the deviation score
COMPUTE MBIT07a = SQRT(MBIT07r*MBIT07r).	Take the square root of the squared deviation of MBIT responses to item 7 to estimate the of absolute value of the deviation score
COMPUTE MBIT09a = SQRT(MBIT09r*MBIT09r).	Take the square root of the squared deviation of MBIT responses to item 9 to estimate the of absolute value of the deviation score
…	Continue with the rest of the items

The ethnocentric biases reflect the difference scores between participant ratings on each of the same items for their ingroup as compared to the targeted outgroup. By taking these difference scores, we thus eliminate any specific biases that might have been associated with the particular content of any item and only extract the contrast between the participant's representation of the targeted outgroup with respect to their own ingroup. Table 3 shows the SPSS syntax for this procedure:

Table 3
SPSS syntax for calculating ethnocentric biases by subtracting the deviations of the responses regarding target outgroup from the deviations of the responses regarding the respondents' ingroup

SPSS Syntax	Mathematical explanation
COMPUTE MBIT01rr = MBIT02r-MBIT01r.	Substract the deviation of the responses about MBIT1 from the responses about MBIT 2
COMPUTE MBIT03rr = MBIT05r-MBIT06r.	Substract the deviation of the responses about MBIT6 from the responses about MBIT 5
COMPUTE MBIT06rr = MBIT11r-MBIT12r.	Substract the deviation of the responses about MBIT12 from the responses about MBIT 11
COMPUTE MBIT09rr = MBIT17r-MBIT18r.	Substract the deviation of the responses about MBIT18 from the responses about MBIT 17
…	Continue with the rest of the items

Manipulation check

For this first step in developing the instrument, it was essential to evaluate whether the accuracy indicators' internal consistencies were acceptable, suggesting that the MBIT scales captured the participants' relative knowledge about targeted groups (e.g., opposite sex or outgroup members). Secondly, it was also important to assess the MBIT internal consistency of the bias indicators scores to determine the degree to which the scales captured the participants' relative implicit biases toward targeted groups.

Before assessing the internal consistency of the bias scores, we applied the Cross-Sample Geometric Mean (CSGM) method (Figueredo et al., 2017) for optimal item selection to assure the cross-cultural comparability in internal consistency of the bias scores. In contrast to an arithmetic mean, characterized for the addition of values, a geometric mean instead multiplies the pertinent parameter estimates (Figueredo et al., 2017). Whereas the arithmetic mean is calculated as a sum divided by n, the geometric mean is computed as the nth root of the product. Through a Boolean lens, a multiplicative term is analogous to a logical "AND" statement (logical conjunction), and an additive term is equivalent to an "OR" statement (logical disjunction). A high-product term is computed if and only if all the multiplied coefficients featured high values (Figueredo et al., 2017). Alternatively, a high additive term can be generated if at least one of the added coefficients is high enough to mitigate any low values in the equation. Consequently, item selection based on a geometric mean instead of an arithmetic mean operates under the logic that the selected items must perform well across all, rather than some, sampled cultures (Figueredo et al., 2017). Unlike other selection methods based on the empirical selection of indicators, this approach does not ignore the sample-specific variances, thus, reducing the capitalization on stochastic fluctuations due to sampling errors. In contrast, items selected based on optimal performance may not replicate well enough when adapting the instrument to a sample outside of the biocultural population in which the scale was initially created (Figueredo et al., 2017).

Most popular methods for empirically selecting items within scales are subject to the problem of capitalization on chance. Item-total correlations are not generally very high for single items, especially as compared to factor loadings on entire multi-item scales in latent variable modeling. It is therefore quite common for such item-total correlations to fall beneath the threshold for statistical significance purely due to random chance.

This problem is especially acute when doing cross-cultural comparisons, as multi-item scales so "tailored" to different cultures (in the mistaken belief by some investigators that they are thereby practicing *cultural* or *historical particularism*), may reflect nothing more than the sampling error that will inevitably exist between the cross-cultural samples.

Because quantitative cross-cultural comparisons require a common scale of measurement, retaining the subset of the items that survived that process of sample-specific selection but are nonetheless still included across all cultures under study is a procedure that is likely to decimate the item pool, particularly since a fair number of those item eliminations will have been unnecessary.

In contrast, the CSGM method does not eliminate any items based on the results obtained from any single culture and does not eliminate any item for one culture that is not eliminated for all others. Instead, this method is designed to select items that performed adequately across *all* of the sampled cultures simultaneously. This procedure is thus somewhat protected against capitalization on chance in item selection decisions, as it is not performed for one cross-cultural sample at a time and repeated independently for each of the other cross-cultural samples.

Finally, it was relevant to determine whether the MBIT bias and accuracy levels remained comparable across targeted groups (local: immigrants; women: men) and across sample cultures (CR and USA). To these end means and variances comparisons within and across samples were performed.

Hypotheses

H1: The factual knowledge of respondents for the opposite sex is higher than for ethnic outgroup members.

H2: The group stereotypes held by respondents for the opposite sex are less homogeneous than those for ethnic outgroup members (Nicaraguan or Mexican Immigrants).

Results

Internal consistency evaluation for accuracy indicators

A slight difference in internal consistencies for the accuracy indicators of the MBIT-Ethnocentrism was identified between USA (α = .73) and Costa Rican (α = .65) samples. Similarly, the USA sample featured a higher internal consistency for accuracy indicators of the MBIT androcentric (α = .80) than the Costa Rican sample (α = .75).

Item selection using cross-sample geometric means

Any *negative item-total correlations*, if of sufficient magnitude, indicate that an item might be measuring the opposite of what the rest of the scale is indicating, and should therefore be eliminated from the scale. Sometimes these negative correlations, however, are of such trivial magnitude that they instead indicate that the affected item may not be measuring anything relevant at all, in which case, it should still be eliminated from the scale.

A total of 8 items exhibiting negative item-total correlations (CR = 3; USA = 5) in the MBIT-Ethnocentrism scale were eliminated from the analyses. Subsequently, a cross-sample geometric mean was computed for each MBIT item correlation across CR and US samples. Seventeen items featuring an average item-total correlation above r = .2 were selected. The analyses with MBIT-Androcentrism values detected 7 negative item-total correlations (CR = 2; USA = 5). Only one item displayed a cross-sample geometric mean across CR and USA samples, larger than r = .2.

Internal consistency evaluation for bias indicators

Subsequent psychometric evaluations with the selected MBIT-Ethnocentrism items (k = 17), identified high internal consistencies for both CR (α = .83) and USA (α = .82) samples. However, as expected from the previous CSGM analysis, the reliability assessment with all MBIT-Androcentrism items (k = 20) detected very low internal consistencies for both CR (α = .34) and USA (α = .26) samples.

Differences within and across cultures

A GLM detected a significant difference in ethnocentrism bias, based on the selected 17 items, between CR and USA samples, with CR participants showing significantly less ethnocentric bias than USA participants. On the other side, a significant difference in androcentric bias, based on the 20 selected items, between CR and USA samples was also identified, but in this case, with CR participants showing significantly more androcentric biases than USA participants (Table 4).

Table 4
Mean scores on ethnocentric bias (locals – immigrants) and androcentric bias (men – women) within and across cultures

	CR		USA			
	M	*SD*	*M*	*SD*	*F-value*	*p-value*
Towards local/immigrants	-0.41	0.54	-0.10	0.56	21.39	< .0001
Towards men/women	0.15	0.66	-0.05	0.48	10.74	< .001

A comparison of accuracy levels of Androcentrism and Ethnocentrism scores did not detect a significant difference within either CR or USA samples, suggesting that participants' objective knowledge of both opposite-sex and outgroup members are relatively comparable (Table 5).

Table 5
Mean scores on accuracy within and across cultures

Mean scores on accuracy $\sqrt{(\text{respondents' answer-correct answer})^2}$						
	Towards local/immigrants		Towards men/women			
	M	*SD*	*M*	*SD*	*F-value*	*p-value*
CR	1.36	.20	1.39	.31	1.02	.451
USA	1.32	.25	1.34	.30	1.01	.464

In contrast, the analyses of variances homogeneity of the accuracies scores detected significant differences within cultures. Furthermore, participants' stereotypes about women across cultures were less homogenous compared to ethnic outgroup members, as indicated by significantly higher variances in accuracy levels for sex than for ethnicity (Table 6).

Table 6
Variance scores on accuracy within and across cultures

Variance scores on accuracy $\sqrt{(\text{respondents' answer-correct answer})^2}$				
	Towards local/immigrants	Towards men/women		
	σ^2	σ^2	*F-value*	*p-value*
CR	.04	.10	2.40	.001
USA	.06	.09	1.44	.011

Discussion

The present study describes the development of a Biased Information Test as a tool for assessing intergroup attitudes. The MBIT follows the principle that the selective perception, processing, and interpretation of social information is associated with intergroup attitudes. This principle is in line with research on other well-documented biases in social cognition as the illusory correlation (Hamilton, 1981), the fundamental attribution error (Ross, 1977), the actor-observer bias (Nisbett et al., 1973), or the ultimate attribution error (Pettigrew, 1979; 2021). The ultimate attribution bias, for instance, is theorized to reflect the expression of ethnocentric bias in the attributions since negative ingroup and positive outgroup outcomes are attributed to situational factors, whereas positive ingroup and negative outgroup outcomes are attributed to causes seen as internal (i.e., stable traits of the groups). Indeed, empirical studies have shown that attributional biases are linked to prejudice (Ensari & Miller, 2005). Therefore, it is reasonable to assess intergroup attitudes from the selective information that people use to describe outgroup members, as the first developers of these tests argued seventy years ago (Campbell, 1950; Hamond, 1948 Loeblowitz-Lennard & Riessman, 1946).

In following this tenet, we generated two-item banks for measuring ethnocentric and androcentric bias. Items provided information about the targeted groups' living conditions, traits, and behaviors (usually a percentage of a proportion), followed by a list of potential answers with the correct option in the middle of the scale and several incorrect options above and below the correct answer. The most challenging facet of developing this type of item is to gather the accurate information of the targeted group from reliable sources (i.e., in the national statistics) and write plausible answers around the actual value. However, once the basic item structure is built, the only pending task is to check for information updates so that participants are exposed to credible data.

Using these tools to measures biases in two cultures, we gather incipient but critical information of their usefulness. The first remarkable result pertains to the internal consistency of the different indicators of the measure. Although the internal consistency for the accuracy indicators was optimal and comparable across cultures, only the internal consistency for ethnocentric biases was adequate across samples, while the internal consistency for androcentric biases was unacceptable in both cultures.

These unexpected low consistencies were not due to a relative lack of knowledge of immigrants compared to women, as there were statistically equivalent mean accuracies for host society members vs. immigrants compared with men vs. women. Instead, data suggest that stereotypes of ethnic outgroup immigrants are more homogeneous than stereotypes of ingroup women.

These results may be produced because male/female relations *within* ethnic groups tend to be more ambivalent than within-group/outgroup relations *between* ethnic groups. According to the Ambivalent Sexism Theory (Glick & Fiske, 1997), the uniqueness of gender relations (i.e., power imbalance and intimate interdependence between men and women) leads intergroup gender attitudes to be profoundly ambivalent based on the coexistence of both hostile and benevolent ideologies toward men and women (Glick & Fiske, 2011). Data support these principles across a great variety of social groups in several cultures (Glick & Fiske, 1996, Glick et al., 2000, 2004). Thus, more research is needed to examine if gender intergroup ambivalence might have affected the MBIT androcentrism's internal consistency.

It is also possible that MBIT androcentrism items induced comparison processes that might trigger ingroup favoritism (Tajfel & Turner, 1979), i.e., women favoring women and men favoring men. Our instrument was developed to measure bias against women, knowing from previous research that both men and women can reproduce the negative cultural stereotypes of women (Glick et al., 2000, 2004; Rollero, Glick, & Tartaglia, 2014); however, this shared cultural knowledge might not have prevented women of evaluating other women more positively than men. There is evidence that when measured by response latencies instruments, women display remarkable strong ingroup preferences, even stronger than men (Leach et al., 2017; Rudman & Goodwin, 2004). This hypothesis should also be adequately tested in future studies, but it is clear from our data so far that our MBIT is not suitable to measure androcentric biases.

Regarding ethnocentric biases, our research shows that Costa Rican participants were generally less ethnocentric toward Nicaraguan Immigrants compared to USA participants toward Mexican immigrants. These results are possibly due to the smaller cultural and ethnolinguistic difference between Costa Rican and Nicaraguans than between North American and Mexican immigrants (Brown & Zagetka, 2011; Turchin, 2003). On the other hand, it must be acknowledged that our samples are composed of university students of two contemporary societies in a historical moment in which the elites of many cultures have converged to become more alike to each other due to the globalization process, and more important, more informed and more politically correct than the rest of the populations of the cultural contexts they belong. Therefore, replications of this line of research with more (within and between cultures) heterogeneous samples will be necessary to provide more insights of the usefulness of these alternative techniques.

In terms of the methodological processes of developing a measure, this study offers important insights. The most important perhaps, was the use of Cross-Sample Geometric Mean (CSGM) method to select items. The CSGM method (Figueredo et al., 2017) is a novel approach to item selection that is designed to identify those items having the highest cross-cultural validity among the samples tested. By using the *geometric* rather than the *arithmetic* mean, his method relies on using the product of the convergent validity coefficients of multiple cross-cultural samples. Whereas the arithmetic mean is inherently *compensatory* as an indicator, permitting a high value in one instance to compensate for a low value in another, the geometric mean instead requires that the convergent validities in all cross-cultural samples be high to achieve a high product term.

Finally, it is important to notice that even our study has important strengths to test the psychometric properties of a novel measure across two different target groups and two different cultures and languages, our study has also some limitations that derive important recommendations for future research. First and foremost, Campbell's (1950) original *Biased Information Test* has gone largely unevaluated in terms of psychometric validity. One key assumption is that the unfavorable *cognitive* biases towards other biocultural groups (historically called *prejudices*) directly tested should be linked to antagonistic *affective* biases towards other groups (historically called *negative ethnocentrism*). In Campbell's reasoning, the negative affective attitudes should presumably have been causal to the negative cognitive biases, thus warranting the cognitive prejudices some plausibility as indicators of the affective animosities. However, this assumption has never been empirically tested, however reasonable it might sound. Thus, an important direction for future research will be somehow examining this relation without the use of flawed methods such as the IAT that purport to give us more direct access to affective biases.

Alternative approaches to validating implicit measures have included examining their correlations to otherwise equivalent explicit measures, and this approach has yielded mixed empirical results (Echaberría Echabe, 2013). However, the logic of doing this seems to us to be logically inconsistent with the entire point of constructing implicit measures to begin with, given that to the extent that implicit and explicit measures correlate, they might be sharing the same self-presentation biases that we are seeking to avoid by constructing the purportedly implicit measures.

None of these more distal objectives was possible to achieve in this initial study, but we have at least been able to update and implement Campbell's (1950) excellent suggestion for a measure of implicit bias and shown some degree of cross-cultural validity. Further, we have been able to circumscribe this validity to the domain of negative ethnocentrism, as opposed to negative androcentrism. This is important because it is just as essential to know for what substantive domains a measure *doesn't* work as it is to know for what domains a measure *does* work. Our results were quite clear in indicating that the MBIT does not work for prejudices presumably stemming from negative androcentrism, or cognitive biases between sexes, but does appear to work quite well for prejudices presumably stemming from negative ethnocentrism, or cognitive biases between biocultural groups.

Author Note. The authors wish to thank Tomás Cabeza de Baca and Paul Robert Gladden for early help in getting this project started.

References

Amodio, D. M., Harmon-Jones, E., & Devine, P. G. (2003). Individual differences in the activation and control of affective race bias as assessed by startle eyeblink response and self-report. *Journal of Personality and Social Psychology, 84*, 738–753. https://doi.org/10.1037/0022-3514.84.4.738

Blanton, H., Jaccard, J., Klick, J., Mellers, B., Mitchell, G., & Tetlock, P. E. (2009). Strong claims and weak evidence: reassessing the predictive validity of the IAT. *The Journal of Applied Psychology, 94*, 567–603. https://doi.org/10.1037/a0014665

Bogardus, E.S. (1925). Measuring social distances. *Journal of Applied Sociology, 9,* 299-308.

Braungart-Rieker, J. M., Hill-Soderlund, A. L., & Karrass, J. (2010). Fear and anger reactivity trajectories from 4 to 16 months: the roles of temperament, regulation, and maternal sensitivity. *Developmental Psychology, 46*, 791-804.

Brown, R., & Zagefka, H. (2011). The dynamics of acculturation: An intergroup perspective. In M. Zanna & J. Olson (Eds). *Advances in Experimental Social Psychology* (Vol. 44, pp. 129-184). Academic Press. https://doi.org/10.1016/b978-0-12-385522-0.00003-2

Brownstein, M., Madva, A., & Gawronski, B. (2020). Understanding Implicit Bias: Putting the Criticism into Perspective. *Pacific Philosophical Quarterly, 101*. https://doi.org/10.1111/papq.12302

Campbell, D. T. (1950). The indirect assessment of social attitudes. *Psychological Bulletin, 47*, 15–38. https://doi.org/10.1037/h0054114

De Houwer, J. (2006). What are implicit measures and why are we using them. In R. W. Wiers & A. W. Stacy (Eds.), *The handbook of implicit cognition and addiction* (pp. 11–28). Thousand Oaks, CA: Sage.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56,* 5–18. https://psycnet.apa.org/doi/10.1037/0022-3514.56.1.5

Dovidio, J. F., Kawakami, K., & Beach, K. R. (2001). Implicit and explicit attitudes: Examination of the relationship between measures of intergroup bias. In R. Brown & S. L. Gaertner (Eds.), *Blackwell handbook of social psychology: Intergroup processes* (pp. 175–197). Blackwell.

Echebarria Echabe, A. (2013). Relationship Between implicit and explicit measures of attitudes: The impact of application conditions. *Europe's Journal of Psychology, 9*, 231-245. https://doi.org/10.5964/ejop.v9i2.544

Ensari, N., & Miller, N. (2005). Prejudice and intergroup attributions: The role of personalization and performance feedback. *Group Processes & Intergroup Relations, 8*, 391–410. https://doi.org/10.1177/1368430205056467

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology, 54*, 297-327. https://doi.org/10.1146/annurev.psych.54.101601.145225

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*, 229–238. https://doi.org/10.1037//0022-3514.50.2.229

Fiedler, K., Messner, C. & Bluemke. M. (2006) Unresolved problems with the "I", the "A", and the "T": A logical and psychometric critique of the Implicit Association Test (IAT), *European Review of Social Psychology,17*, 74 147, https://doi.org/10.1080/10463280600681248

Figueredo, A. J., Garcia, R. A., Menke, J. M., Jacobs, W. J., Gladden, P. R., Bianchi, J., ... & Li, N. P. (2017). The K-SF-42: A new short form of the Arizona Life History Battery. *Evolutionary Psychology, 15*, https://doi.org /10.1177/1474704916676276.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132*, 692–731. https://doi.org/10.1037/0033-2909.132.5.692

Gawronski, B. (2019). Six Lessons for a Cogent Science of Implicit Bias and Its Criticism. *Perspectives on Psychological Science, 14*, 574–595. https://doi.org/10.1177/1745691619826015

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*, 300–312. https://doi.org/10.1177/0146167216684131

Glick, P. & Fiske, S. T. (1996). The Ambivalent Sexism Inventory: Differentiating hostile and benevolent Sexism. *Journal of Personality and Social Psychology, 70,* 491-512. https://doi.org/10.1037/0022-3514.70.3.491

Glick, P., & Fiske, S. T. (1997). Hostile and benevolent Sexism: Measuring ambivalent sexist attitudes toward women. *Psychology of Women Quarterly, Special Issue: Measuring attitudes toward appropriate roles for men and women, 21,* 119-135. https://doi.org/10.1111/j.1471-6402.1997.tb00104.x

Glick, P., & Fiske, S. T. (1999). The Ambivalence toward Men Inventory: Differentiating hostile and benevolent beliefs about men. *Psychology of Women Quarterly, 23,* 519-536. https://doi.org/10.1111/j.1471-6402.1999.tb00379.x

Glick, P., & Fiske, S. T. (2011). Ambivalent sexism revisited. *Psychology of Women Quarterly, 35,* 530-535. https://doi.org/10.1177/ 0361684311414832

Glick, P., Fiske, S. T., Mladinic, A., Saiz, J. L., Abrams, D., ... López López, W. (2000). Beyond prejudice as simple antipathy: hostile and benevolent sexism across cultures. *Journal of Personality and Social Psychology, 79,* 763–775. https://doi.org/10.1037//0022-3514.79.5.763

Glick, P., Lameiras, M., Fiske, S. T., Eckes, T., Masser, B., Volpato, C., Manganelli, A. M., Pek, J., Huang, L., Sakalli-Ugurlu, N., Castro, Y. R., D'Avila Pereira, M. L., Willemsen, T. M., Brunner, A., Six-Materna, I, & Wells, R. (2004). Bad but bold: Ambivalent attitudes toward men predict gender inequality in 16 nations. *Journal of Personality and Social Psychology, 86,* 713–728. https://doi.org/10.1037/0022-3514.86.5.713

Goodall, C. E. (2011). An overview of implicit measures of attitudes: Methods, mechanisms, strengths, and limitations. *Communication Methods and Measures, 5,* 203–222. https://doi.org/10.1080/19312458.2011.596992

Greenwald, A. G. Banaji, M. R. Rudman, L. A. Farnham, S. D. Nosek, B. A. Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review,* 109, 3–25. https://doi.apa.org/ 10.1037/0033-295X.109.1.3

Greenwald, A., Brendl, M., Cai, H., Charlesworth, T.E., Dovidio, J., ... Sherman, J. (2020). The Implicit Association Test at age 20: What is known and what is not known about implicit bias. (2020, April 7). *PsyArXiv*. https://doi.org/10.31234/osf.io/bf97c

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. https://doi.org/10.1037/0022-3514.74.6.1464

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*, 139–150. https://doi.org/10.2307/2086306

Hamilton, D. L. (1981). Illusory correlation as a basis for stereotyping. In Hamilton, D. L. (Ed.), *Cognitive processes in stereotyping and intergroup behavior* (pp. 115-144). Erlbaum.

Hammond, K. R. (1948). Measuring attitudes by error-choice; an indirect method. *Journal of Abnormal Psychology, 43*, 38–48. https://doi.org/10.1037/h0059576

Harris, L. T., & Fiske, S. T. (2007). Social groups that elicit disgust are differentially processed in mPFC. *Social Cognitive and Affective Neuroscience*, *2*, 45–51. https://doi.org/10.1093/scan/nsl037

Hefner, D., Rothmund, T., Klimmt, C., & Gollwitzer, M. (2011). Implicit measures and media effects research: Challenges and opportunities. *Communication Methods and Measures, 5*, 181–202. https://doi.org/10.1080/19312458.2011.597006

Heitland, K., & Bohner, G. (2011). An Indirect Paper-and-Pencil Measure of Prejudice: A German Version of the Racial Argument Scale. *Current Psychology, 30*, 1–21. https://doi.org/10.1007/s12144-010-9098-2

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*, 1369–1385. https://doi.org/10.1177/0146167205275613

Jacoby, L. L., Debner, J. A. & Hay, J. F. (2001). Proactive interference, accessibility bias, and process dissociations: Valid subject reports of memory. *Journal of Experimental Psychology: Memory, Learning, and Cognition*, 27, 686–700. https://doi.org/10.1037/0278-7393.27.3.686

Koopman, J., Howe, M., Johnson, R. E., Tan, J. A., & Chang, C.-H. (2013). A framework for developing word fragment completion tasks. *Human Resource Management Review*, *23*, 242–253. https://doi.org/10.1016/ j.hrmr.2012.12.005

Lai, C. K., & Wilson, M. E. (2021). Measuring implicit intergroup biases. *Social and Personality Psychology Compass, 15*, Article e12573. https://doi.org/10.1111/spc3.12573

Leach, C. W., Carraro, L., Garcia, R. L., & Kang, J. J. (2017). Morality stereotyping as a basis of women's in-group favoritism: An implicit approach. *Group Processes & Intergroup Relations, 20*, 153–172. https://doi.org/10.1177/1368430215603462

Loeblowitz-Lennard, H., & Riessman, F., Jr. (1946). A proposed projective attitude test. *Psychiatry: Journal for the Study of Interpersonal Processes, 9*, 67–68. https://doi.org/10.1080/00332747.1946.11022583

Mann, N. H., & Kawakami, K. (2012). The long, steep path to equality: Progressing on egalitarian goals. *Journal of Experimental Psychology: General, 141*, 187–197. https://doi.org/10.1037/a0025602

Matsuda, Y. T., Okanoya, K., & Myowa-Yamakoshi, M. (2013). Shyness in early infancy: approach-avoidance conflicts in temperament and hypersensitivity to eyes during initial gazes to faces. *PloS one, 8*, e65476.

Meissner F, Grigutsch LA, Koranyi N, Müller F and Rothermund K (2019) Predicting Behavior with Implicit Measures: Disillusioning Findings, Reasonable Explanations, and Sophisticated Solutions. *Frontiers in Psychology, 10,* 2483. https://doi.org/10.3389/fpsyg.2019.02483

Moisés, A. (2012). Sexualidad en Mesoamérica: machismo y marianismo. *Científica,* 1, 45-53. http://hdl.handle.net/11715/349

Nisbett, R. E., Caputo, C., Legant, P., & Marecek, J. (1973). Behavior as seen by the actor and as seen by the observer. *Journal of Personality and Social Psychology, 27*, 154–164. https://doi.org/10.1037/h0034779

Olson, M. A., & Zabel, K. L. (2016). Measures of prejudice. In T. D. Nelson (Ed.), *Handbook of prejudice, stereotyping, and discrimination* (p. 175–211). Psychology Press.

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*, 171–192. https://doi.org/10.1037/a003273

Park, R.E. (1924). The concept of social distance as applied to the study of racial attitudes and racial relations. *Journal of Applied Sociology, 8,* 339-344.

Pettigrew, T. F. (1979). The Ultimate attribution error: Extending Allport's cognitive analysis of prejudice. *Personality and Social Psychology Bulletin, 5*, 461–476. https://doi.org/10.1177/014616727900500407

Pettigrew, T. F. (2021). *Contextual social psychology: Reanalyzing prejudice, voting, and intergroup contact*. American Psychological Association. https://doi.org/10.1037/0000210-000

Rae, J. R., & Olson, K. R. (2018). Test–retest reliability and predictive validity of the Implicit Association Test in children. *Developmental Psychology, 54,* 308–330. https://doi.org/10.1037/dev0000437

Rollero, C., Glick, P., & Tartaglia, S. (2014). Psychometric properties of short versions of the Ambivalent Sexism Inventory and Ambivalence Toward Men Inventory. *TPM: Testing, Psychometrics, Methodology in Applied Psychology, 21*, 149–159. https://doi.org/10.1037/t33821-000

Ross, L. (1977). The Intuitive psychologist and his shortcomings: Distortions in the attribution process. In *Advances in experimental social psychology* (Vol. 10, pp. 173-220). Academic Press.

Rudman, L. A., & Goodwin, S. A. (2004). Gender differences in automatic in-group bias: why do women like women more than men like men? *Journal of Personality and Social Psychology, 87*, 494–509. https://doi.org/10.1037/0022-3514.87.4.494

Smith-Castro, V. & Molina, M. (2011). *La entrevista cognitiva: Guía para su aplicación en la evaluación y mejoramiento de instrumentos de papel y lápiz*. Serie Cuadernos Metodológicos [The cognitive interview: Guide for its application in the evaluation and improvement of paper and pencil instruments]. Instituto de Investigaciones Psicológicas.

Stevens, E. (1973). Marianismo The Other Face of Machismo. In A. Pescatello (Ed.) *Female and male in Latin America* (p. 89–101). Pittsburgh University Press.
Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.). *The Social psychology of intergroup relations* (pp. 33–48). Brooks-Cole.
Torching, P. (2003). *Historical dynamics*. Princeton, NJ: Princeton University Press.
Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.
van Ravenzwaaij, D., van der Maas, H. L. J., & Wagenmakers, E.-J. (2011). Does the name-race Implicit Association Test measure racial prejudice? *Experimental Psychology, 58*, 271–277. https://doi.org/10.1027/1618-3169/a000093
Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12,* 129–140. https://doi.org/10.1080/17470216008416717
Webb, T. L. (2011). Advice-taking as an unobtrusive measure of prejudice. *Behavior Research Methods, 43*, 953–963. https://doi.org/10.3758/s13428-011-0122-8
Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Sage Publications.