An Empirical Comparison of Four Generalized Trust Scales: Test-Retest Reliability, Measurement Invariance, Predictive Validity, and Replicability*

Blaine G. Robbins

New York University Abu Dhabi

September 18, 2021

Forthcoming at *Sociological Methods & Research* (Version 2)

Abstract

The Stranger Face Trust scale (SFT) and Imaginary Stranger Trust scale (IST) are two new self-report measures of generalized trust that assess trust in strangers—both real and imaginary—across four trust domains. Prior research has established the reliability and validity of SFT and IST, but a number of measurement validation tests remain. Across three separate studies, I assess the test-retest reliability, measurement invariance, predictive validity, and replicability of SFT and IST, with the misanthropy scale (MST) and generalized social trust scale (GST) serving as benchmarks. First, tests of internal consistency, test-retest reliability, and longitudinal measurement invariance established that all four generalized trust scales were acceptably reliable, with SFT and IST yielding greater overall reliability than MST and GST. Second, tests of multiple group measurement invariance revealed that SFT and IST were equivalent across gender, race, education, and age-groups, while MST and GST were non-equivalent across the same sociodemographic groups. Third, an investment game established the predictive validity of SFT and MST, with IST and GST yielding poor predictive validity. Fourth, tests of factor structure and measurement invariance indicated that all four generalized trust scales replicated across samples. The present findings bolster the validity, reliability, and measurement equivalence of SFT and IST, while illustrating the compromised validity and measurement non-equivalence of MST and GST. Implications for the measurement of generalized trust are discussed.

Keywords: Generalized trust, reliability, validity, measurement invariance

Corresponding author: Blaine Robbins, New York University Abu Dhabi, Building A5 1191, PO Box 129188, Abu Dhabi, United Arab Emirates

Email: bgr3@nyu.edu

* I am indebted to Maria Grigoryeva, Rebecca Morton, and Georgia Michailidou for comments, conversations, and/or suggestions. This research was conducted, in part, at the Social Science Experimental Laboratory (SSEL) at New York University Abu Dhabi in February 2019. I would also like to thank Cian Farrelly and Kerstin Brünig from Qualtrics as well as Alizeh Batra and Arus Hakhnazaryan from SSEL for their invaluable research assistance. An earlier version of this manuscript was presented at the 2019 Annual Meeting of the American Sociological Association.

Introduction

A number of prominent social scientists claim that generalized trust—or the optimism and unconditional faith people have in unknown others—is necessary for large-scale social order (Almond and Verba 1963; Fukuyama 1995; Knack and Keefer 1997; Putnam 1993, 2000; Sztompka 1999; Uslaner 2002; Zak and Knack 2001). When levels of generalized trust are low, countries experience economic stagnation (Algan and Cahuc 2013; Tabellini 2010), stunted political participation (Almond and Verba 1963; Putnam 2000), rampant political corruption (La Porta et al. 1997; Bjørnskov 2010; Robbins 2012a; Rothstein 2011; Uslaner 2008a), and elevated rates of interpersonal violence (Messner et al. 2004; Rosenfeld et al. 2001). Conversely, when civil society, economic equality, informal social ties, religious communities, and political institutions are robust, generalized trust tends to flourish (Berggren and Jordahl 2006; Bjørnskov 2007; Brehm and Rahn 1997; Delhey and Newton 2005; Dinesen 2012ab, 2013; Dinesen and Hooghe 2010; Fairbrother and Martin 2013; Freitag and Buhlmann 2009; Glanville et al. 2013; Herreros 2004; Paxton 2002, 2007; Robbins 2011, 2012b; Rothstein and Uslaner 2005; Sønderskov and Dinesen 2016; Stolle 1998; Traunmüller 2011; Tsai et al. 2011; Uslaner 2002; Welch et al. 2004; You 2012).

To date, a handful of survey instruments have informed the bulk of this research. The classic and most frequently used measure is the *most-people trust question* (Rosenberg 1956): “Some people say that most people can be trusted. Others say you can’t be too careful in your dealings with people. How do you feel about it?” Since the instrument’s inception, efforts have been made to alter the question’s wording, structure, scale length, and dimensionality (Bauer and Freitag 2018; Couch and Jones 1997; Lundmark et al. 2016; Uslaner 2011, 2015). Yet, the reliability and validity of these measures has come under increasing empirical scrutiny (Delhey

et al. 2011; Glanville and Paxton 2007; Miller and Mitamura 2003; Reeskens and Hooghe 2008; Robbins 2019; Sturgis and Smith 2010; Torpe and Lolle 2011; van Hoorn 2014), raising doubts about the associations and trends observed in the literature as well as the scientific community's ability to make comparisons across studies, countries, and time. If unrectified, the continued use of these measures will hinder scientific progress and the accumulation of knowledge.

In an effort to move the literature forward, Robbins (2019) recently proposed two new self-report measures of generalized trust: a Stranger Face Trust scale (SFT) and an Imaginary Stranger Trust scale (IST). The first scale, SFT, shows respondents the faces of six strangers drawn from the Chicago Face Database (Ma et al. 2015), and asks respondents to assess each stranger along four matters for which trust is placed. The faces vary according to gender and race, while the matters concern pecuniary and non-pecuniary behaviors of general interest to individuals. The second scale, IST, asks respondent to imagine meeting a total stranger for the first time and identify the extent to which they would trust said imaginary stranger along the four matters previously mentioned. SFT was created as a long-form questionnaire, while IST was fashioned as a short-form questionnaire for easy implementation by the research community.

These newly proposed scales were evaluated in terms of face validity and content validity, and subject to various measurement validation tests, vis-à-vis common generalized trust scales (Robbins 2019). With respect to face validity and content validity, Robbins (2019) claims—like many others (Bauer and Freitag 2018; Colquitt et al. 2007; Erikson 1950; Rotter 1967)—that the concept of generalized trust “... is best thought of as a baseline starting point at which people (dis)trust strangers for any number of matters (p. 7).” Because prior measures of generalized trust do not operationalize specific strangers *and* particular matters for which trust is placed, such measures have low face validity and content validity. With SFT and IST, Robbins

(2019) treats “...generalized trust as a latent construct in which observed indicators of generalized trust manifest as assessments of trustworthiness for *particular matters* (in the case of SFT and IST) regarding *specific strangers* (in the case of SFT) (p. 8).” He contends, as a result, that SFT and IST have greater face validity and content validity than common measures of generalized trust. Regarding measurement validation, Robbins (2019) uses confirmatory factor analysis and structural equation models to establish the factorial validity, internal consistency, convergent validity, discriminant validity, concurrent validity, and known-groups validity of SFT and IST. The results also indicate that common generalized trust scales—the Misanthropy scale (MST) and Generalized Social Trust scale (GST)—may be less valid and less reliable than SFT and IST (Robbins 2019).¹

Despite initial empirical support for the validity and internal consistency of SFT and IST, some tests of reliability and validity remain unexplored. First, while a large literature has shown that common measures of generalized trust are fairly stable across time (Baumert et al. 2017; Dawson 2019; Giordano et al. 2012; Kiley and Vaisey 2020; Uslander 2002), the test-retest reliability of SFT and IST has not been established. Identifying the extent to which trends in generalized trust arise from instability of measurement over time (versus true temporal patterns and changes) is important to demonstrate for any measure. Second, establishing measurement invariance has recently become a central requirement of measurement validation. Measurement invariance, or measurement equivalence (MI/E), is the degree to which the measurement model of a latent construct is the same across groups and time (Liu et al. 2017; Putnick and Bornstein 2016; Vandenberg 2002; Vandenberg and Lance 2000). While research has investigated the

¹ MST consists of the dichotomous most people trust, most people fair, and most people helpful questions. GST consists of the 4-point trust people you meet for the first time, people of another religion, and people of another nationality questions.

cross-cultural MI/E of common generalized trust measures (Reeskens and Hooghe 2008; Freitag and Bauer 2013), research has yet to explore whether the MI/E of SFT, IST, and other measures of generalized trust holds across sociodemographic groups, samples, and time. Without establishing MI/E, observed differences between groups and time points may be an artifact of the measurement instrument and not necessarily true differences in levels of generalized trust.

Third, the criterion validity of common generalized trust measures is well known (Bauer et al. 2019; Paxton and Glanville 2015; Robbins 2016a, 2019), especially their ability to predict behavior in economic games (Fehr et al. 2002; Gächter et al. 2004; Lo Iacono 2018; Peysakhovich et al. 2014; Sapienza et al. 2013). Although Robbins (2019) reveals that SFT and IST exhibit strong concurrent validity, the extent to which SFT/IST predict cooperation remains unknown. Predictive validity is one of the most powerful tools for establishing a measure's overall validity. If a survey item fails to predict a concrete behavioral outcome that is representative of the latent construct of interest, then the survey item may be of little utility. For instance, if standardized tests and driving exams are valid, then they should be predictive of college grades and driving records; the same goes for the relation between SFT/IST and investments in a trust game. Fourth, even though replicability is not an explicit goal of measurement validation, it is nonetheless critical for measurement validation tests to yield consistent results across groups and time (Freese and Peterson 2017). For instance, a robust set of measurement tools should yield similar factor structures, internal consistencies, and levels of convergent validity and discriminant validity in samples A through Z.

The goals of the present manuscript are to assess the test-retest reliability, MI/E, predictive validity, and replicability of SFT, IST, MST, and GST. To achieve these goals, I employed data gathered from three separate studies. Study 1 examines data used in Robbins

(2019), which consisted of 2,041 Qualtrics web-panel members. Data from study 1 were used as a benchmark for establishing the factor structure of all latent constructs, to assess MI/E across groups and studies, and to investigate the replicability of all four generalized trust scales. In study 2, data came from a longitudinal Qualtrics web-panel ($N = 1,751$) in which measurement validation tests were executed across two time-points (a six-week gap). The goal of study 2 was to establish test-retest reliability and MI/E across groups, studies, and time, and to explore replicability. In study 3, data came from a behavioral economic laboratory subject pool ($N = 146$), which allowed me to assess the predictive validity of the four generalized trust scales (an eight-week gap), and to further probe MI/E and replicability.

The three studies yield several important findings. First, tests of internal consistency, test-retest reliability, and longitudinal MI/E show that all four scales are acceptably reliable measures of generalized trust, but that SFT/IST appear to be more reliable than MST/GST. Second, tests of multiple group MI/E reveal that the factor loadings and item intercepts of SFT/IST were invariant and equivalent across gender, race, education, and age-groups, while the factor loadings and item intercepts of MST/GST were variant and non-equivalent—to varying degrees—across the same sociodemographic groups. Third, results from an investment game show that SFT and MST were moderately (weakly) associated with behavioral trust (trustworthiness), while IST and GST were both weakly associated with behavioral trust and trustworthiness. Fourth, the factor structures observed in study 1—standardized factor loadings, average variances extracted, intercorrelations between factors—and assessments of convergent validity and discriminant validity were replicated in studies 2 and 3. Multiple group MI/E also show that the factor loadings and item intercepts of all four generalized trust scales were invariant and equivalent across the three studies. In sum, the present findings further bolster the validity, reliability, and

measurement invariance of SFT and IST, while illustrating the compromised validity and measurement non-equivalence of MST and GST. Implications for the measurement of generalized trust are reviewed in the discussion and conclusion.

Data and Methods

To investigate the reliability, validity, and measurement invariance of SFT and IST, three separate studies were conducted. Note that throughout the manuscript I will refer to, and use interchangeably, (a) factor as latent construct, and (b) item as observed indicator.

Study 1

Participants. Participants were 2,041 members of a Qualtrics market research web-panel from the original Robbins (2019) study, which is based on a general population non-probability quota sample of adults in the United States. In terms of relative majorities, 50.4 percent of the participants were female, 36.7 percent had a high school diploma or equivalent, 81.3 percent were non-Hispanic white, and 29.9 percent were between the ages of 18 and 34 ($M = 47.2$, $SD = 16.62$, $\min = 18$, $\max = 99$). For more information about sociodemographic characteristics, see Robbins (2019).

Procedure. The Qualtrics survey consisted of 14 blocks: some of which were anchored at the beginning of the study (i.e., consent form, screener question, and sociodemographic characteristics), while the remaining blocks were shown in random order from respondent-to-respondent. For the present study, I use blocks reserved for the sociodemographic characteristics and measures of generalized trust (see Robbins 2019). Participants completed the study over a two-week period in April of 2018. Eligibility was restricted to U.S. adults age 18 and older. Participants who completed the study received an incentive, where the type of compensation varied (e.g., cash, airline miles, etc.). The median survey length was approximately 27 minutes.

Measurement of generalized trust: SFT and IST. This paper examines two recently developed scales of generalized trust (Robbins 2019): SFT and IST. SFT consists of six faces drawn from the Chicago Face Database (Ma et al. 201), which is a data base that consists of 597 high-resolution photographs of male and female human faces of varying age and ethnicity (<https://chicagofaces.org/>). The six faces were selected based on four criteria: age (median U.S. ages), race (Caucasian, African American, and Latino/a), gender (male and female), and perceived trustworthiness (neutral evaluations) normed by an independent rater sample. For more information about each face, see Robbins (2019) and Ma et al. (2015). Headshots of all six human faces can be found in the Supplemental Materials online.

For each human face, respondents were asked to “imagine meeting the following stranger for the first time. Please identify how much you would trust this stranger for each of the following.” For IST, respondents were not shown human faces. Instead, respondents were asked to “imagine meeting a total stranger for the first time. Please identify how much you would trust this stranger for each of the following.” SFT and IST contained four domains (or matters) of trust for which respondents would rate each of the six human faces (SFT) and the imaginary stranger (IST): (a) “To keep a secret that is damaging to your reputation” (SECRET); (b) “To repay a loan of one thousand dollars” (LOAN); (c) “To look after a child, family member, or loved one while you are away” (CHILD); and (d) “To provide advice about how best to manage your money” (ADVICE).

Each trust domain was measured using a 4-point rating scale. The rating scale anchored on *Do not trust at all* and *Trust completely*, with *Do not trust very much* and *Trust somewhat* in-between the anchors and a *Don’t know* option at the end of the scale. For SFT, the order of the six human faces was randomized between respondents. The order of the four trust domains was

also randomized between human faces. For IST, only the order of the four domains was randomized between respondents.

Measurement of generalized trust: MST and GST. Two instruments common to the General Social Survey and World Values Survey were used as benchmark comparisons for SFT/IST: a 3-item Misanthropy scale (Brehm and Rahn 1997; Glanville et al. 2013; Paxton 1999; Mewes et al. 2021) and a 3-item Generalized Social Trust scale (Delhey et al. 2011; Newton and Zmerli 2011; Robbins 2016). MST consists of 3 items scored on a dichotomous answer scale. Item 1 (TRUST) asks “Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?” with *most people can be trusted* and *need to be very careful in dealing with people* as anchors with a *don’t know* option. Item 2 (FAIR) asks “Do you think that most people would try to take advantage of you if they got the chance or would they try to be fair?” with *most people would try to take advantage of me* and *most people try to be fair* as anchors with a *don’t know* option. Item 3 (HELP) asks “Would you say that most of the time people try to be helpful or that they are mostly looking out for themselves?” with *people mostly try to be helpful* and *people mostly look out for themselves* as anchors with a *don’t know* option. *Don’t know* responses were treated as system missing and all 3 items were recoded to parallel polarity (i.e., 1 = most people can be trusted, most people try to be fair, and people mostly try to be helpful).²

² To be consistent with prior research (Brehm and Rahn 1997; Glanville et al. 2013; Paxton 1999; Mewes et al. 2021), Robbins (2019) and I measure MST with dichotomous answer scales. One drawback of this methodological approach is that shorter answer scales produce less valid measurements than 7- or 11-point answer scales, and decrease the detail in which latent factors can be measured (Krosnick and Fabrigar 1997; Lundmark et al. 2016). As a result, our findings may not extend to other MST scales, such as those generated by the European Social Survey (see Zmerli and Newton 2008; Reeskens and Hooghe 2008). An important avenue of future research would replicate the analyses presented here, and in Robbins (2019), but with indicators of MST using longer answer scales. I thank an anonymous reviewer for raising this important point.

GST consists of 3 items scored on a 4-point rating scale anchoring at *Do not trust at all* and *Trust completely* with *Do not trust very much* and *Trust somewhat* in-between the anchors and a *Don't know* option at the end of the scale. The question asked “Could you tell me for each whether you trust people from this group completely, somewhat, not very much, or not at all?” with grid items being “people you meet for the first time” (FIRST), “people of another religion” (RELI), and “people of another nationality” (NATION). *Don't know* responses were treated as system missing.

Study 2

Study 2 aims to replicate tests of internal consistency, factor structures, and MI/E found in study 1 and to additionally assess test-retest reliability and longitudinal MI/E of SFT, IST, MST, and GST.

Participants. For this study, members of a Qualtrics market research longitudinal web-panel were recruited to participate, which yielded a general population non-probability quota sample of 1,751 U.S. adults. In relation to study 1, 51.6 percent of the participants were female, 25 percent had a high school diploma or equivalent, 83.4 percent were non-Hispanic white, and 25.9 percent were between the ages of 18 and 34 ($M = 49.7$, $SD = 16.01$, min = 19, max = 91). For more information about sociodemographic characteristics, see the Supplemental Materials online.

Procedure. The Qualtrics survey used a block structure similar to study 1, but with a reduced number of thematic blocks. For this study, blocks contained the consent form, a screener question, and survey items of sociodemographic characteristics and trust. Compensation, data collection, and eligibility mirrored the procedures outlined in study 1. Participants completed the study two times at an interval of 45 days between the months of January and March of 2019.

1,751 participants completed the survey at time 1, while 1,066 participants completed the survey at time 2. The median survey length at time 1 and time 2 was approximately 17 minutes and 15 minutes, respectively.

Measurement. SFT, IST, MST, and GST were measured as described in study 1.

Study 3

Study 3 aims to replicate the tests of internal consistency and factor structures found in studies 1 and 2, and to additionally assess the predictive validity and MI/E of SFT, IST, MST, and GST across samples.

Participants. 146 participants of a behavioral economic laboratory subject pool were recruited to participate in study 3. 43 percent of the participants were male with a mean age of 20.9 years ($SD = 1.59$, min = 18, max = 29). The three largest racial and ethnic groups were white (30.8 percent), east Asian (16.4 percent), and south Asian (16.4 percent). The majority of the participants identified “no religion” as their religious preference, while the next three largest religious groups were Muslim (15.7 percent), Orthodox-Christian (12.8 percent), and Roman Catholic (10 percent).

Procedures. All undergraduate student participants of the laboratory subject pool (~1000 students) were sent recruitment emails via Hroot with a Qualtrics survey link. The survey used a block structure similar to studies 1 and 2. Blocks contained the consent form and survey items of sociodemographic characteristics, trust, and other psychological scales. Participants who completed the survey received an incentive of approximately 8USD in local currency. The data collection stopping rule consisted of two weeks of recruitment with a reminder email sent during the middle and toward the end of data collection. Eligibility was restricted to adults age 18 and older who were undergraduate student participants of the laboratory subject pool and who

voluntarily consented to participate. Participants completed the survey over a two-week period in February of 2019. The median survey length was approximately 35 minutes.

After 60 days, the 146 participants who completed the survey were sent recruitment emails via Hroot and asked to participate in a laboratory experiment. In total, 88 undergraduate students participated in the experiment.³ The experiment was coded in zTree v.4.1.0, and took about 30 minutes for participants to complete. Subjects were paid based on an exchange rate of approximately 2 tokens for 0.27USD (in local currency) plus a show-up fee of 8.17USD (in local currency). The experiments transpired in 8 separate sessions over 3 days in April of 2019.

After arrival to the laboratory, participants sat at a cubicle of their choice. Once everyone was seated, participants provided their informed consent, read instructions, and were asked to complete a few paper-based exercises to confirm their understanding of the game. Participants then played one round of the classic investment (or trust) game (Berg et al. 1995). At the start of the experiment, participants were randomly assigned to one of two roles: player A or player B. Participants were informed of these two roles, and were told that they would be randomly matched to another participant in the room. Both players started the experiment with 100 tokens. Participants were told that player A moved first, and that player A had the opportunity to transfer all, some, or none of their 100 tokens to player B. Each token player A sent to player B was tripled. If player A sent 50 tokens to player B, for example, player B would receive 150 tokens. The amount player A transferred to player B would be displayed on both players' screens. Player

³ One participant took the experiment twice. The data for this participant's second experimental session—along with data of the randomly assigned partner—were removed from the analysis. For two of the experimental sessions, there were an odd number of participants. In these sessions, one of the odd numbered participants was randomly assigned to player A, while the experimenter assumed the role of player B. The experimenter always returned a fixed proportion of tokens to player B: 50 percent of the transferred amount. Data generated by the two odd-numbered subjects were retained for the analysis, while data generated by the experimenter were excluded. Due to how the laboratory was physically structured, participants were not aware of the experimenter's participation.

B then had the possibility to return tokens to player A. Player B could choose any amount between zero and the tripled amount player A transferred to player B. The amount player B returned to player A would not be tripled. The 100 tokens player B started the game with could not be sent to player A. All participants were informed of and aware of these rules. The experiment ended after player B submitted the amount of tokens he or she would like to return to player A.

Measurement. SFT, IST, MST, and GST were measured as described in study 1. Trust in the investment game was measured as the proportion of tokens sent to player B. Trustworthiness in the investment game was measured as the proportion of tokens returned to player A. Trust measured in the investment game was used as a criterion variable for the validation of SFT, IST, MST, and GST. Trustworthiness measured in the investment game was used to assess discriminant validity. The logic behind this distinction is as follows. Trust and trustworthiness are two different constructs. Trust is often conceptualized as person A's belief about the competence and motivation of person B with respect to matter Y (i.e., what is of interest to person A), while trustworthiness is person B's actual level of competence and motivation to achieve matter Y on behalf of person A (Bauer 2013; Cook et al. 2005; Hardin 2002; Robbins 2016b). In spite of the theoretical distinctions between trust and trustworthiness, they may be correlated empirically. Person A, for instance, might fall prey to a false consensus effect and anchor on their own trustworthiness in deciding whether to trust person B (Butler et al. 2015); that is, person A may ask themselves "How would I behave as person B under these circumstances?", which would inform person A's trust in person B. A valid self-report measure of generalized trust, then, should predict the proportion of tokens sent in an investment game

(behavioral trust), and may predict—to a small degree—the proportion of tokens returned (behavioral trustworthiness).

Data Analysis

Reliability

Internal consistency was estimated with Cronbach's α . Cronbach's α values range between 0 and 1, with values greater than .70 indicating acceptable reliability (Nunnally 1978). Test-retest reliability was assessed by the correlation between factors at a 45-day interval. Similar to Cronbach's α , correlations between factors greater than .70 indicate acceptable test-retest reliability. Notably, test-retest reliability assumes that respondents' true scores do not change from the first measurement to the second measurement. A moderate correlation, therefore, might be indicative of an unreliable test, of true changes in the persons being measured, or both. To separate reliability of measurement from (in)stability of a true score, I employ longitudinal MI/E (Liu et al. 2017), which I review later on.

Factor structure

Robbins (2019) found that all latent constructs under study (SFT, IST, MST, GST) were best represented by a single factor. For SFT, in particular, a higher-order CFA in which trust domain indicators (SECRET, LOAN, CHILD, ADVICE) loaded on first-order factors for each stranger's face—white female (WF), white male (WM), black female (BF), black male (BM), Latin female (LF), Latin male (LM)—which in turn loaded on a second-order factor (SFT), yielded model fit similar to a single-factor model. In the single-factor model, the four trust domains (SECRET, LOAN, CHILD, ADVICE) of each face (WF, WM, BF, BM, LF, LM) were summed and divided by four while ignoring missing values (i.e., row means). Each of the 4-item scales for WF, WM, BF, BM, LF, and LM were treated as observed continuous indicators of a

single-factor SFT latent construct. Building on these findings, I treat SFT as a 6-item single factor in which SFT loads on WF, WM, BF, BM, LF, and LM in the present manuscript.

To test the four-factor structure of SFT, IST, MST, and GST, robust full-information maximum likelihood estimation (MLR) was performed using Mplus 8.1 (Muthén and Muthén 2017). Since CFAs estimated with MLR are relatively robust against deviations from the assumption of normality (DiStefano 2002; Muthén and Kaplan 1985), I model all indicators as continuous variables. Even so, I present results of CFAs estimated with mean-and-variance adjusted weighted least squares (WLSMV) in the Supplemental Materials online, which is an appropriate estimation technique for modeling categorical (binary and ordinal) indicators (Flora and Curran 2004; Lei 2009; Lubke and Muthén 2004). The CFAs using MLR for continuous indicators and WLSMV for categorical indicators are substantively similar. Finally, under the assumption of data missing completely at random or missing at random, MLR for missing data provides asymptotically unbiased estimates (Allison 2001). I have no reason to believe that the data are missing not at random given the non-sensitive nature of the questions. As a result, I use MLR to model missingness throughout.

To investigate the convergent and discriminant validity of the four-factor structure of SFT, IST, MST, and GST, I calculated a number of measurement validation statistics common to classical test theory. For convergent validity of observed indicators, standardized factor loadings greater than .70 suggest acceptable convergent validity since more than 50 percent (i.e., $.70^2$) of the variance in an indicator is explained by the latent construct to which the indicator is theoretically related (Campbell and Fiske 1959; Kline 2005). I also assess the convergent validity of the measurement model by calculating the average variance extracted (AVE). AVE measures the level of variance captured by a latent construct versus the level due to measurement error

(Fornell and Larcker 1981). The AVE for a latent construct is calculated by summing the variances extracted (i.e., R^2) from each indicator over the total number of indicators. AVE values greater than .70 are considered acceptable. To establish the degree to which latent constructs converge or discriminate from one another, I evaluate the correlations between latent constructs. Correlations less than .85 indicate discriminant validity, while correlations greater than .85 indicate that two latent constructs overlap and likely measure the same latent concept. I also evaluate discriminant validity by comparing the AVE of a latent construct to the shared variance with other constructs. To establish discriminant validity, the AVE for each latent construct should be greater than the squared correlation between latent constructs (Fornell and Larcker 1981).

To assess model fit, several indices of fit were used, including the root mean squared error of approximation (RMSEA), the comparative fit index (CFI), and the standardized root mean square residual (SRMR). The RMSEA is a sample-free index that measures the difference between the estimated model and the population covariance matrix. RMSEA ranges between 0 and 1, with values less than .08 indicating adequate model fit (Hu and Bentler 1999). The CFI, in contrast, adjusts for issues of sample size inherent to the χ^2 test of model fit, and investigates the difference between the data and the estimated model. CFI ranges from 0 to 1, with values greater than .90 suggesting adequate model fit (Hu and Bentler 1999). The SRMR is a standardized measure of the square root of the difference between the sample covariance matrix and the model covariance matrix. Like the RMSEA, SRMR ranges between 0 and 1, with values less than .08 indicating adequate model fit (Muthén and Muthén 2017).

Measurement invariance/equivalence

MI/E across gender (men, women), race (white, non-white), education (HSD or less, some college, college degree, graduate degree), age-groups (18-34, 35-49, 50-64, 65+), samples (study 1, study 2, study 3), and time-points (study 2 at time 1, study 2 at time 2) was examined by testing and comparing three nested models with multiple group analysis (Vandenberg and Lance 2000; Liu et al. 2017; Putnick and Bornstein 2016). In multiple group analysis, each successive model includes the previous model restrictions plus additional constraints. Model 1, the configural invariance model, tests the equivalence of the factor structure. Model 2, the metric invariance model, tests the equivalence of the factor loadings. Model 3, the scalar invariance model, tests the equivalence of measurement intercepts. The nested models are tested with χ^2 difference tests, but since the test is sensitive to sample size I rely on changes to the CFI, RMSEA, and SRMR across models. Following Cheung and Rensvold (2002) and Chen (2007), I use criteria of a .01 change in CFI paired with changes in RMSEA and SRMR of .015. If nested model comparisons reject the χ^2 difference test, yield $\Delta CFI > .01$, $\Delta RMSEA > .015$, and $\Delta SRMR > .015$, then the null hypothesis of invariance should be rejected.

Predictive validity

To test the predictive validity of SFT, IST, MST, and GST, two statistics were used. First, correlations between measures of generalized trust with the proportion of tokens sent in the investment game were computed. Second, measures of generalized trust were correlated with the proportion of tokens returned in the investment game. A valid measure of generalized trust will be moderately-to-strongly correlated with the proportion of tokens sent but negligibly-to-weakly correlated with the proportion of tokens returned.

Results

Study 1

Descriptive statistics and reliability. Item means, standard deviations, number of observations, and Cronbach α are shown in Table 1. For SFT and IST, Cronbach's α ranged from .929 to .962, indicating good internal consistency; the four trust domains of each face (e.g., WF) also exhibited good internal consistency, with Cronbach's α between .920 and .930. In contrast, the Cronbach α 's for MST ($\alpha = .686$) and GST ($\alpha = .741$) yielded reasonable internal consistency.

Factor structure. In order to identify the CFA, I constrained the first unstandardized factor loading of each factor to 1. Factor variances as well as intercepts and residuals variances of the observed indicators were freely estimated and statistically significant (not shown but available upon request). Results are provided in Table 2. The CFA for study 1 indicated that the four-factor CFA provided a good fit to the data, $\chi^2(98) = 651.310$, RMSEA = .053, CFI .960, and SRMR = .070. For SFT and IST, all standardized factor loadings were well above the standard cutoff value of .70 with most factor loadings greater than .90. The AVEs for SFT (.806) and IST (.763) were also well above the standard cutoff of .70. In contrast, some standardized factor loadings were below .70 for MST and GST, while others were above .70. Likewise, the AVEs for MST and GST were below the acceptable cutoff of .70. Overall, the convergent validity of observed indicators of SFT/IST was superior to that of MST/GST.

With respect to discriminant validity, the intercorrelations between latent constructs were all less than .85, while the AVEs for each latent construct were greater than the squared correlations between latent constructs. This set of findings reveals latent construct discrimination between SFT, IST, MST, and GST, but with some conceptual overlap between latent constructs. The correlations between SFT and IST ($r = .714$) and MST and GST ($r = .600$) were greater than all other correlations between latent constructs, indicating that SFT and IST have more

conceptual overlap with each other than with MST or GST (which, in turn, exhibit more conceptual overlap with each other than with SFT or IST). Overall, the findings indicate discriminant validity between latent constructs.

The present findings—using old data with a different estimation technique—support the “robustness” (Freese and Peterson 2017) of Robbins’ (2019) results, which were originally estimated, via WLSMV, as a six-factor CFA with categorical observed indicators. Four-factor CFAs in which observed indicators were treated as categorical can be found in the Supplemental Materials online (the results were substantively similar, but the standardized factor loadings and AVEs were larger for categorical than continuous indicators).

Measurement invariance. The analysis of MI/E across gender, race, education, and age-groups for SFT, IST, MST, and GST are reported in Tables 3, 4, 5, and 6, respectively. In each table, I report nested model comparisons between models that estimated configural invariance, metric invariance, and scalar invariance by group (i.e., gender, race, education, and age-group).⁴ Starting with SFT in Table 3, many of the χ^2 difference tests were statistically significant but none of the changes to SRMR, RMSEA, and CFI indicated that the null hypothesis of MI/E should be rejected. In other words, SFT’s factor loadings and item intercepts were invariant by gender, race, education, and age-group. Results for IST are reported in Table 4. Like SFT, many of the χ^2 difference tests were statistically significant but, again, none of the changes to SRMR,

⁴ To identify the configural invariance model, I follow Vandenberg (2002) and Vandenberg and Lance (2000) and constrained the factor means and variances to 0 and 1, respectively, across groups. Factor loadings as well as item intercepts and residuals variances were freely estimated across groups. To identify the metric invariance model, I constrained the factor loadings to equality across groups, constrained the factor variance to 1 for the first group but freely estimated the factor variances for all other groups, constrained the factor means to 0 across groups, and freely estimated item intercepts and residual variances across groups. To identify the scalar invariance model, I constrained the factor loadings and item intercepts to equality across groups, constrained the factor variance and mean to 1 and 0, respectively, for the first group but freely estimated the factor variances and means for all other groups, and freely estimated item residual variances across groups.

RMSEA, and CFI indicated that the null hypothesis of MI/E should be rejected. Note, however, that IST may exhibit metric non-equivalence across age-groups as three of the four tests rejected the null hypothesis of MI/E (yet $\Delta CFI < .01$).

For MST, Table 5 shows that factor loadings were non-equivalent by gender, education, and age-group. This means that factor loadings differ significantly between men and women, across levels of education, and between older and younger respondents. Item intercepts were non-equivalent across age-groups, which indicates that MST's item intercepts differ between older and younger respondents. In Table 6, tests of MI/E for GST showed that factor loadings were invariant by gender, race, education, and age-groups, but that item intercepts were non-equivalent by gender, education, and age-groups. In other words, GST's item intercepts differ significantly between men and women, across levels of education, and between older and younger respondents.

In short, SFT/IST appear to have better MI/E by gender, race, education, and age-groups than MST/GST. A set of findings reproduced for models in which items were estimated as categorical indicators or when SFT was estimated as a higher-order factor (see the Supplemental Materials online).

Study 2

Descriptive statistics and reliability. Descriptive statistics in Table 1 indicate comparable item means and standard deviations between study 1 and study 2. Cronbach's α ranged from .678 to .968 with values greater for SFT/IST than MST/GST, indicating poor-to-acceptable internal consistency depending on the factor (similar levels and patterns of internal consistency were observed in study 1).

Assessments of test-retest reliability reveal that the correlations between factors was .604 for SFT ($p < .001$) and .526 for IST ($p < .001$), indicating poor test-retest reliability.⁵ In contrast, the correlations between factors was .892 for MST ($p < .001$) and .677 for GST ($p < .001$), indicating acceptable-to-good test-retest reliability. Overall, results of test-retest reliability show that MST/GST exhibit greater stability across time than SFT/IST.

Factor structure. Results are reported in Table 2 (study 2 column). Using the identification strategy outlined in study 1, the four-factor CFA estimated in study 2 largely replicated the factor structure outlined in study 1. The CFA provided a good fit to the data ($\chi^2(98) = 445.427$; RMSEA = .045; CFI = .969; SRMR = .068) and yielded standardized factor loadings, AVEs, and correlations between factors similar to those found in study 1. Like study 1, the convergent validity of observed indicators of SFT/IST was superior to that of MST/GST, and the results indicate discriminant validity between latent constructs.

Four-factor CFAs in which observed indicators were treated as categorical can be found in the Supplemental Materials online. The Supplemental Materials online also contain a replication of the higher-order CFA estimated by Robbins (2019) in which observed indicators of SFT were treated as categorical. The higher-order CFA estimated in study 2 replicated the original higher-order CFA found in Robbins (2019).

Measurement invariance. Using the identification strategies outlined in study 1, results of MI/E by gender, race, education, and age-groups are reported in the Supplemental Materials online. For SFT and IST, the findings are comparable to study 1, while MST and GST are

⁵ To assess test-retest reliability, four two-factor CFAs were estimated in which factors covaried across time, e.g., $\text{Corr}(\text{SFT1}, \text{SFT2})$. For each model, residuals variances freely covaried across time, and MLR was used to adjust for panel attrition (39.1 percent dropout from time 1 to time 2). In order to identify the CFAs, the first factor loading of each factor was constrained to 1. See the Supplemental Materials online for CFAs estimated with and without a missing data technique adjusting for panel attrition.

slightly more invariant across sociodemographic groups in study 2 than study 1. In spite of these differences, SFT and IST still exhibit better MI/E by gender, race, education, and age-groups than MST and GST. This set of findings reproduced for models in which items were estimated as categorical indicators.

With respect to longitudinal MI/E, the identification strategy of configural, metric, and scalar invariance models is similar to multiple group MI/E except for two key differences: factors and residual variances freely covary across time. Findings indicate that factor loadings and item intercepts for SFT, IST, MST, and GST are time invariant (see the Supplemental Materials online). In other words, an equality of the factor loadings and item intercepts was observed across time for the four latent constructs under study. The results of the longitudinal MI/E—when coupled with assessments of test-retest reliability—indicate that changes to SFT and IST scores can be attributed to true changes in the latent constructs. SFT and IST, as a result, are reliable measures of generalized trust. The same conclusion can be drawn for MST and GST. These findings were robust to estimation strategies in which indicators are treated as categorical or panel attrition is modeled with listwise deletion (see the Supplemental Materials online).

Study 3

Descriptive statistics and reliability. Descriptive statistics in Table 1 indicate comparable item means and standard deviations across the three studies. In study 3, Cronbach's α ranged from .675 to .954 with values greater for SFT/IST than MST/GST, indicating poor-to-good internal consistency depending on the factor. While similar levels and patterns of internal consistency were observed in study 1 and study 2, Cronbach's α was slightly smaller for the four trust domains of each face (i.e., WF). Yet, the Cronbach's α for SFT (i.e., WF, WM, BF, BM, LF, LM) was .954, which is comparable to SFT's α in study 1 and study 2.

Descriptive statistics for the investment game are as follows. Player As, on average, sent 33 percent of their tokens to player Bs ($N = 45$, $M = .333$, $SD = .322$, $\min = 0$, $\max = 1$). Nine players sent 0 tokens, while six players sent 100 percent of their tokens. Player Bs, on average, returned 31.6 percent of their tokens to player As ($N = 43$, $M = .316$, $SD = .253$, $\min = 0$, $\max = .777$).⁶ Ten players returned 0 tokens, while no players returned 100 percent of their sent tokens. The average payoff was approximately 26USD in local currency (Player A = 21.78USD; Player B = 30.84USD), which consisted of the monetary units earned in the investment game plus the show-up fee.

Factor structure. Results are reported in Table 2 (study 3 column). Using the identification strategy outlined in study 1, the four-factor CFA estimated in study 3 replicated the factor structures outlined in study 1 and study 2. The CFA provided a good fit to the data ($\chi^2 (98) = 185.776$; $RMSEA = .078$; $CFI = .921$; $SRMR = .075$) and yielded standardized factor loadings, AVEs, and correlations between factors comparable to those found in study 1 and study 2. Like the previous studies, the convergent validity of observed indicators of SFT/IST was superior to that of MST/GST, and the results indicate discriminant validity between latent constructs.

Measurement invariance. Because of sample size restrictions, it was not possible to assess MI/E by gender, race, education, and age-groups. MI/E, however, was tested across samples. Using the identification strategies outlined in study 1, factor loadings and item intercepts for SFT, IST, MST, and GST appear to be invariant (see the Supplemental Materials online).⁷ The results suggest that there is an equality of the factor structure across samples for all

⁶ For the analysis of the proportion of tokens returned, the overall $N = 35$. Recall that nine players sent 0 tokens, one of which sent 0 tokens to an experimenter (in the role of player B) whose data were excluded from the analysis.

⁷ For MST, item intercepts are equivalent across samples only when indicators are treated as categorical.

four generalized trust scales. This set of findings supports the “repeatability” (Freese and Peterson 2017) of SFT, IST, MST, and GST.⁸

Predictive validity. Results are reported in Figures 1 and 2. Each panel in Figure 1 (Figure 2) is a scatterplot of the relation between the proportion of tokens sent (returned) in the investment game and a latent factor score generated from a single-factor CFA of SFT (Panel A), IST (Panel B), MST (Panel C), and GST (Panel D). Best fit linear regression lines are red, 95% CIs are grey dashed lines, and individual observations are blue circles. Given attrition observed from the survey ($N = 146$) to the experiment ($N = 88$), response propensity weighting was used for all estimates and figures (Gelman 2007; Guo and Fraser 2015).⁹

Figure 1 shows that all four generalized trust scales were positively associated with the proportion of tokens sent in the investment game, but that SFT ($r = .390$) and MST ($r = .262$) were the only statistically significant scales at the $p < .05$ level. The positive correlation between IST and the proportion of tokens sent was weak ($r = .134$, $p = .394$), while the correlation coefficient for GST was close to zero ($r = .053$, $p = .736$). Figure 2, in contrast, reveals that the associations between the four generalized trust scales and the proportion of tokens returned in the investment game were all statistically non-significant at the $p < .05$ level. Although statistically non-significant, differences in magnitude reveal that the association is stronger for $\text{MST} > \text{IST} >$

⁸ The undergraduate student body sampled in study 3 is one of the most diverse in the world. Recent graduating classes consist of students representing over 80 nationalities and 60 languages. While not the strongest test of cross-cultural MI/E, the results provisionally suggest that SFT and IST’s factor structures are invariant across national origin. Large-scale research endeavors sampling Western and Non-Western nations, however, are required to systematically determine the cross-cultural equivalence of SFT and IST.

⁹ For each respondent, a response propensity score was estimated based on a selected model, then weighting adjustments were set to the inverse of these response propensities. The selected logit model consisted of regressing the propensity to participate in the laboratory experiment (0 = survey only, 1 = survey + experiment) on variables with no item nonresponse (i.e., complete observations) measured at the time of the survey. These variables included the latent factor scores for SFT, IST, MST, and GST, as well as measures of social value orientations, risk taking, betrayal aversion, prior trusting behavior, gender, religious denomination, years of education, and associational membership.

GST > SFT. For instance, the correlation coefficient for MST was weak ($r = .191, p = .242$), while the correlation coefficient for SFT was close to zero ($r = -.034, p = .811$). This set of findings is robust to alternative modeling procedures (see the Supplemental Materials online) and potential outliers.¹⁰ Taken together, SFT and MST exhibit acceptable predictive validity, while IST and GST display poor predictive validity.

Discussion and Conclusion

The analysis of generalized trust has a long history in the social sciences, both as a cause and as a consequence of various individual and societal outcomes (Fukuyama 1995; Knack and Keefer 1997; Putnam 1993, 2000; Uslaner 2002; Zak and Knack 2001). Appropriate methods for the measurement of generalized trust are needed to draw unbiased and valid conclusions. Yet, available tools used to measure generalized trust, such as the *most-people trust* question and the misanthropy scale, lack face validity and content validity: specific strangers *and* particular matters for which trust is placed are not operationalized with common measures of generalized trust (Bauer and Freitag 2018; Cook et al. 2005; Hardin 2002, 2006). This ultimately creates a problem of measurement non-equivalence, where survey questions and answer scales are interpreted differently across respondents (Delhey et al. 2011; Reeskens and Hooghe 2008; Uslaner 2002; Sturgis and Smith 2010; Torpe and Lolle 2011). To address these outstanding measurement problems, Robbins (2019) recently developed two new measures of generalized trust—the Stranger Face Trust scale and the Imaginary Stranger Trust scale—and showed that both were more valid and more internally consistent than MST and GST across a number of measurement validation tests.

¹⁰ For models predicting the proportion of tokens sent or returned, the observations produced (a) CooksD values below the cutoff value of $4/N$, (b) approximately normal deviance residuals, and (c) small diagonal “hat” matrix and studentized residuals. This suggests that influence *and* leverage are not biasing the estimates.

In this article, I further investigated the reliability and validity of old (MST and GST) and new (SFT and IST) measures of generalized trust, examined their measurement invariance across groups, samples, and time, and explored their replicability. Tests of measurement validation revealed the following. In terms of reliability, I found that the internal consistency of observed indicators was good for SFT and IST, and acceptable for MST and GST. Moreover, the current findings indicate that levels of internal consistency for SFT, IST, MST, and GST replicate across samples. Assessments of test-retest reliability and longitudinal MI/E yielded similar conclusions. While the correlations between factors were much more stable for MST/GST than SFT/IST, longitudinal MI/E revealed that the factor loadings and item intercepts for SFT, IST, MST, and GST were time invariant. Thus, all four scales yield stable measurement over time. SFT and IST, however, do exhibit change in true scores, which suggests that generalized trust may be less stable and more amenable to change than previously thought. Taken together, the results indicate that SFT, IST, MST, and GST are reliable measures of generalized trust, but that the reliability of SFT/IST is greater than that of MST/GST (given the results of internal consistency, test-retest reliability, *and* longitudinal MI/E).

In terms of MI/E across groups, I discovered that SFT and IST were invariant by gender, race, education, and age-groups. The factor loadings and item intercepts of SFT and IST were the same between men and women, whites and non-whites, college educated and non-college educated, as well as older and younger subpopulations. This equivalence was shown in study 1 and then repeated in study 2. MST and GST, in contrast, appeared to be more non-equivalent across groups. Depending on the sample and scale, the factor loadings and/or item intercepts of MST and GST differed by gender, race, education, and age-groups. The amount of non-equivalence by groups, however, did vary from sample-to-sample. Overall, the findings indicate

a high degree of measurement invariance across classic sociodemographic groups for SFT and IST, and a low degree of measurement invariance for MST and GST.

With regard to MI/E across samples, all four generalized trust scales yielded measurement invariance between studies. What this means is that SFT, IST, MST, and GST produced equivalent factor loadings and item intercepts from sample-to-sample for each respective scale. Confirmatory factor analysis further supports this conclusion: standardized factor loadings, AVEs, and intercorrelations between factors were comparable across studies, which yielded consistent patterns in the convergent validity and discriminant validity of SFT, IST, MST, and GST. The convergent validity of SFT/IST, for instance, was always superior to that of MST/GST regardless of the sample, with greater discriminant validity between SFT/IST and MST/GST than between SFT and IST (or MST and GST). This evidence, when taken together, indicates that the factor structures of the four generalized trust scales were repeatable and replicable across samples (Freese and Peterson 2017).

With respect to predictive validity, SFT/MST and IST/GST yielded substantively different results. SFT and MST were moderately correlated with behavioral trust measured in an investment game (statistically significant associations), but weakly correlated with behavioral trustworthiness (statistically non-significant associations). IST and GST, by contrast, were both weakly correlated with behavioral trust *and* trustworthiness (statistically non-significant associations).

Implications and Directions for Future Research

In conjunction with Robbins' (2019) original paper, a number of important implications and directions for future research can be drawn from the current set of studies. All four generalized trust scales yield surprisingly consistent results across samples and time. This is an

important finding that supports the replicability of the reliabilities and factor structures under study. SFT/IST and MST/GST do, however, diverge in a number of important respects.

First, MST/GST consistently yield poorer convergent validity than SFT/IST, and the degree of convergence *and* divergence between all four latent constructs is stable across samples. These results support Robbins' (2019) original finding that SFT and IST capture different underlying constructs than MST and GST, and that the latent constructs of SFT and IST are better measured by their observed indicators than MST or GST.

Second, the analysis shows that the MI/E of MST and GST is compromised across classic sociodemographic groups. This implies that prior differences observed in generalized trust between known-groups (e.g., education) might be an artifact of the measurement instrument used (MST and GST) and not necessarily true, unbiased differences in levels of generalized trust (see Alesina and La Ferrara 2002; Freitag and Bühlmann 2009; Paxton 2007; Robbins 2011; Simpson 2006). Importantly, this biased feature of measurement escapes SFT and IST: both scales measure generalized trust similarly across sociodemographic groups (e.g., between college and non-college educated respondents).

Third, SFT/IST are less stable across time than MST/GST, and this instability does not appear to be a function of measurement. Moreover, SFT/IST are equivalent across age groups, while MST/GST are non-equivalent. This implies that generalized trust—as measured by SFT and IST—is more susceptible to change than originally assumed (Baumert et al. 2017; Dawson 2019; Kiley and Vaisey 2020; Uslaner 2002). If generalized trust is relatively changeable, then it might be open to environmental influences and personal experiences throughout the life course (Glanville and Paxton 2007; Paxton and Glanville 2015), and less a persistent, asituational human trait or psychological predisposition inherited early on in life via cultural transmission

(Uslaner 2002, 2008b; Wu 2020). Addressing these questions with the use of SFT and IST is an important direction for future research.

The results also indicate that well known trends in generalized trust may, in reality, be a function of differences in measurement across ages and cohorts. A common finding is that generalized trust is on the decline in the United States (Clark and Eisenstein 2013; Mewes et al. 2021; Paxton 1999; Putnam 2000; Robinson and Jackson 2001; Schwadel and Stout 2012; Wilkes 2011). Some scholars attribute this decline to cohort replacement, with younger, less trusting cohorts replacing older, more trusting ones (Putnam 2000; Robinson and Jackson 2001). Robbins (2019) finds that MST/GST are positively correlated with age, while SFT/IST are negatively correlated with age. And, in the present article, I find that MST/GST are non-equivalent across age groups, while SFT/IST are equivalent across age groups. Together, these results suggest that age differences in how individuals interpret MST and GST could be driving the observed declines in generalized trust, and that arguments in favor of cohort replacement should be reevaluated. The extent to which these changes are a function of true declines and/or systematic measurement error is an important question for future research.

Fourth, a valid measure of generalized trust should converge upon criterion measures of trust *but* diverge from measures of trustworthiness. The two criterion measures I use to assess predictive validity are the proportion of tokens sent (behavioral trust) and the proportion of tokens returned (behavioral trustworthiness) in an investment game (Berg et al. 1995). Contrary to expectations, the literature indicates that the *most-people trust* question and other generalized trust scales correlate with behavioral trust (Lo Iacono 2018; Fehr et al. 2002; Naef and Schupp 2009; Sapienza et al. 2013) *and* trustworthiness (Ashraf et al. 2006; Glaeser et al. 2000; Sapienza et al. 2013). In the present article, SFT and MST meet the criteria laid out for predictive validity:

both are statistically significant and moderately correlated with behavioral trust, but statistically non-significant and weakly correlated with behavioral trustworthiness. IST and GST, on the other hand, fail to meet one of the necessary criteria: both are weakly correlated with behavioral trust. In short, SFT and MST converge upon standard conceptions of generalized trust by meeting the expected criterion-related validity of (a) an association between survey measure and behavioral trust, and (b) a negligible association between survey measure and behavioral trustworthiness.

In Table 7, I have compiled what we know so far from the current set of studies and Robbins (2019) about the reliability, validity, and measurement invariance of all four generalized trust scales. The overall takeaway is that SFT and IST are superior measures of generalized trust: across a number of measurement validation tests, they yield greater—or in some cases equivalent—reliability, validity, and measurement invariance than MST and GST. Beyond assessing relative performance of the two sets of measures, the current results demonstrate that MST and GST, along with their observed indicators, introduce unnecessary bias into statistical estimates. Because of this, it is difficult for researchers to identify whether scientific discoveries using MST and GST are—or will be—a function of true patterns and relations or an artifact of measurement. Rectifying these issues requires a shift in the protocols used to measure generalized trust, such as the widespread introduction of SFT and IST. That being said, because indicators of MST and GST are available in a number of large-scale population surveys (e.g., General Social Survey, World Values Survey, etc.), their use in social science research will undoubtedly continue for years to come. I recommend that researchers carefully choose between MST and GST (as well as their respective indicators) based on the findings shown in the present article and in Robbins (2019). Even though MST and GST perform well across a number of

measurement validation tests, I would like to make a general call for more transparency in published research, where scholars discuss issues of measurement underlying MST and GST (e.g., discriminant validity, measurement non-equivalence across sociodemographic groups, etc.), and what these shortcomings imply for empirical findings.

Given their strengths, are there any downsides to SFT or IST? The first downside is time. SFT takes approximately 5 minutes for respondents to complete. To minimize bias, it also requires the order of faces to be randomly assigned, and the order of trust domains to be randomly assigned to each face. These are, rightfully so, limiting factors. But if the survey mode permits, I would argue that, in light of the magnitude of gains to reliability, validity, and measurement invariance, SFT should be used whenever generalized trust is treated as a key independent *or* dependent variable (regardless of the length of the survey). If the survey mode is a constraining feature (e.g., telephone survey), or if generalized trust is a mere control variable, then my suggestion is to use IST over MST or GST, since IST is a more valid and reliable measure of generalized trust than common alternatives.

The second downside is representation. MST and GST are established measures of generalized trust that have been used in representative random samples for decades across a multitude of cultures and societies. The representation of SFT and IST has yet to be established to such a degree. Doing so would require a major research program. If pursued, the first step in this program would be to establish the reliability and validity of SFT and IST with representative random samples from countries across the globe. This is an important endeavor given the non-equivalence of common measures of generalized trust both within *and* between regions of the world (Delhey et al. 2011; Lolle and Torpe 2011; Reeskens and Hooghe 2008). Currently, it is an outstanding question whether—and to what degree—SFT and IST are valid outside of Western

contexts. Relatedly, future tests should broaden the racial diversity of the faces used in SFT. The current manifestation of SFT does not include Asian faces. With the inclusion of East Asian and South Asian racial-ethnic groups, the faces used would encompass the majority of racial groups observed in the world, which may strengthen the cross-cultural equivalence of SFT. At the same time, future research must consider alternative trust domains. The current rating scales are not exhaustive and may benefit from the inclusion of low-cost everyday matters for which people trust others, such as trusting a stranger to return a lost wallet or parcel. Similarly, it would be worthwhile to investigate how the number of scale points impacts the overall reliability and validity of SFT and IST, and whether employing longer rating scales for indicators of MST and GST produce greater convergence between SFT/IST and MST/GST. It may be the case that divergence in measurement between all four generalized trust scales derives, in part, from the length of answer scales, where longer rating scales yield greater convergence (see Lundmark et al. 2016). While two critical pieces of research have demonstrated the value of SFT and IST, much more work is required.

Conclusion

After reviewing the survey-based and experimental research on trust, Russell Hardin (2006) wrote that “...there is relatively little to learn about trust from these two massive research programs. Without retuning their protocols to address standard conceptions of trust, they cannot contribute much to understanding trust as we generally know it, and they cannot play a very constructive role in explaining social behavior, institutions, or social and political change (p. 74).” My hope, with this article, was to further contribute to the “retuning of protocols” first called for by Hardin (2006) and executed by Robbins (2019).

Author’s Note

Data and code have been deposited in the Open Science Framework
(<https://osf.io/sx9af/>).

References

- Alesina, Alberto and Eliana La Ferrara. 2002. "Who Trusts Others?" *Journal of Public Economics* 85: 207-234.
- Algan, Yann and Pierre Cahuc. 2013. "Trust and Growth." *Annual Review of Economics* 5: 521-49.
- Allison, Paul D. 2001. *Missing Data*. London, UK: Sage Publications.
- Almond, Gabriel A. and Sidney Verba. 1963. *The Civic Culture*. Princeton, NJ: Princeton University Press.
- Ashraf, Nava, Iris Bohnet, and Nikita Piankov. 2006. "Decomposing Trust and Trustworthiness." *Experimental Economics* 9: 193-208.
- Bauer, Paul. 2013. "Conceptualizing Trust and Trustworthiness." *Political Concepts Working Paper Series, No. 61*.
- Bauer, Paul and Markus Freitag. 2018. "Measuring Trust." In Eric Uslaner, ed., *Oxford Handbook of Social and Political Trust*, pp. 15-36. Cambridge, UK: Oxford University Press.
- Bauer, Paul C., Florian Keusch, and Frauke Kreuter. 2019. "Trust and Cooperative Behavior: Evidence from the Realm of Data-Sharing." *PLoS ONE* 14: e0220115.
- Baumert, Anna, Anna Halmburger, Tobias Rothmund, and Christian Schemer. 2017. "Everyday Dynamics in Generalized Social and Political Trust." *Journal of Research in Personality* 69: 44-54.
- Berg, Joyce, John Dickhaut, and Kevin McCabe. 1995. "Trust, Reciprocity, and Social History." *Games and Economic Behavior* 10: 122-142.
- Berggren, Niclas and Henrik Jordahl. 2006. "Free to Trust: Economic Freedom and Social Capital." *Kyklos* 59: 141-169.
- Bjørnskov, Christian. 2007. "Determinants of Generalized Trust: A Cross-Country Comparison." *Public Choice* 130: 1-21.
- Bjørnskov, Christian. 2010. "How Does Social Trust Lead to Better Governance? An Attempt to Separate Electoral and Bureaucratic Mechanisms." *Public Choice* 144: 323-346.
- Brehm, John and Wendy Rahn. 1997. "Individual-Level Evidence for the Causes and Consequences of Social Capital." *American Journal of Political Science* 41: 999-1023.
- Butler, Jeffrey V., Paola Giuliano, and Luigi Guiso. 2015. "Trust, Values, and False Consensus." *International Economic Review* 56: 889-915.
- Campbell, Donald T. and Donald W. Fiske. 1959. "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." *Psychological Bulletin* 56: 81-105.
- Chen, Fang Fang. 2007. "Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance." *Structural Equation Modeling* 14: 464-504.
- Cheung, Gordon W. and Roger B. Rensvold. 2002. "Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance." *Structural Equation Modeling* 9: 233-255.
- Clark, April K. and Marie A. Eisenstein. 2013. "Interpersonal Trust: An Age-Period-Cohort Analysis Revisited." *Social Science Research* 42: 361-375.
- Colquitt, Jason A., Brent A. Scott, and Jeffery A. LePine. 2007. "Trust, Trustworthiness, and Trust Propensity: A Meta-Analytic Test of their Unique Relationships with Risk Taking and Job Performance." *Journal of Applied Psychology* 92: 909-27.
- Cook, Karen S., Russell Hardin, and Margaret Levi. 2005. *Cooperation without Trust?* New York, NY: Russell Sage.
- Couch, Laurie L. and Warren H. Jones. 1997. "Measuring Levels of Trust." *Journal of Research in Personality* 31: 319-336.

- Dawson, Chris. 2019. "How Persistent is Generalized Trust?" *Sociology* 53: 590-599.
- Delhey, Jan and Kenneth Newton. 2005. "Predicting Cross-National Levels of Social Trust: Global Pattern or Nordic Exceptionalism?" *European Sociological Review* 21: 311-327.
- Delhey, Jan, Kenneth Newton, and Christian Welzel. 2011. "How General Is Trust in 'Most People'? Solving the Radius of Trust Problem." *American Sociological Review* 76:786-807.
- Dinesen, Peter. 2012a. "Does Generalized (dis)trust Travel? Examining the Impact of Cultural Heritage and Destination-Country Environment on Trust of Immigrants." *Political Psychology* 33: 495-511.
- Dinesen, Peter. 2012b. "Parental Transmission of Trust or Perceptions of Institutional Fairness: Generalized Trust of Non-Western Immigrants in a High-Trust Society." *Comparative Politics* 44: 273-289.
- Dinesen, Peter T. 2013. "Where You Come from or Where You Live? Examining the Cultural and Institutional Explanation of Generalized Trust Using Migration as a Natural Experiment." *European Sociological Review* 29: 114-128.
- Dinesen, Peter and Marc Hooghe. 2010. "When in Rome, Do as the Romans Do: The Acculturation of Generalized Trust Among Immigrants in Western Europe." *International Migration Review* 44: 697-727.
- DiStefano, Christine. 2002. "The Impact of Categorization with Confirmatory Factor Analysis." *Structural Equation Modeling* 9: 327-346.
- Erikson, Erik. H. 1950. "Growth and Crisis of the 'Healthy Personality'." In M. J. E. Senn, ed., *Symposium on the Health Personality*, pp. 91-146. Oxford, UK: Josiah Macy, Jr. Foundation.
- Fairbrother, Malcolm and Isaac W. Martin, 2013. "Does Inequality Erode Social Trust? Results from Multilevel Models of US States and Counties." *Social Science Research* 42: 347-360.
- Fehr, Ernst, Urs Fischbacher, Bernhard Von Rosenblatt, Jürgen Schupp, and Gert Wagner. 2002. "A Nationwide Laboratory Examining Trust and Trustworthiness by Integrating Behavioural Experiments into Representative Surveys." *Schmollers Jahrbuch* 122: 519-542.
- Flora, D. B. and P. J. Curran. 2004. "An Empirical Evaluation of Alternative Methods of Estimation for Confirmatory Factor Analysis with Ordinal Data." *Psychological Methods* 9: 466-491.
- Fornell, Claes and David F. Larcker. 1981. "Evaluating Structural Equation Models with Unobservable Variables and Measurement Error." *Journal of Marketing Research* 18: 39-50.
- Freese, Jeremy and David Peterson. 2017. "Replication in Social Science." *Annual Review of Sociology* 43: 147-165.
- Freitag, Markus and Paul Bauer. 2013. "Testing for Measurement Equivalence in Surveys: Dimensions of Social Trust across Cultural Contexts." *Public Opinion Quarterly* 77: 24-44.
- Freitag, Markus and Marc Bühlmann. 2009. "Crafting Trust: The Role of Political Institutions in a Comparative Perspective." *Comparative Political Studies* 42: 1537-1566.
- Freitag, Markus and Richard Traummüller. 2009. "Spheres of Trust: An Empirical Analysis of the Foundations of Particularized and Generalized Trust." *European Journal of Political Research* 48: 782-803.
- Fukuyama, Francis. 1995. *Trust: The Social Virtues and the Creation of Prosperity*. New York: Free Press.
- Gächter, Simon, Benedikt Herrmann, and Christian Thöni. 2004. "Trust, Voluntary Cooperation, and Socio-Economic Background: Survey and Experimental Evidence." *Journal of Economic Behavior & Organization* 55: 505-531.

- Gelman, Andrew. 2007. "Struggles with Survey Weighting and Regression Modeling." *Statistical Science* 22: 153-164.
- Giordano, Giuseppe Nicola, Jonas Björk, and Martin Lindström. 2012. "Social Capital and Self-Rated Health—A Study of Temporal (Causal) Relationships." *Social Science & Medicine* 75: 340-348.
- Glaeser, Edward L., David I. Laibson, Jose A. Scheinkman, and Christine L. Soutter. 2000. "Measuring Trust." *The Quarterly Journal of Economics* 115: 811-846.
- Glanville, Jennifer L. and Pamela Paxton. 2007. "How do we Learn to Trust? A Confirmatory Tetrad Analysis of the Sources of Generalized Trust." *Social Psychology Quarterly* 70: 230-242.
- Glanville, Jennifer L., Matthew A. Andersson, and Pamela Paxton. 2013. "Do Social Connections Create Trust? An Examination using New Longitudinal Data." *Social Forces* 92: 545-562.
- Guo, Shenyang and Mark W. Fraser. 2015. *Propensity Score Analysis: Statistical Methods and Applications, 2nd Edition*. New York, NY: Sage.
- Hardin, Russell. 2002. *Trust & Trustworthiness*. New York: Russell Sage.
- Hardin, Russell. 2006. *Trust*. Cambridge, UK: Polity.
- Herreros, Francisco. 2004. *The Problems of Forming Social Capital. Why Trust?* New York: Palgrave.
- Hu, L. and P. M. Bentler. 1999. "Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives." *Structural Equation Modeling* 6: 1-55.
- Kiley, Kevin and Stephen Vaisey. 2020. "Measuring Stability and Change in Personal Culture Using Panel Data." *American Sociological Review*.
- Kline, Rex B. 2005. *Principles and Practice of Structural Equation Modeling (2nd ed.)*. New York, NY: Guilford Press.
- Knack, Stephen and Philip Keefer. 1997. "Does Social Capital Have an Economic Payoff? A Cross-Country Investigation." *The Quarterly Journal of Economics* 112: 1251-1288.
- Krosnick, Jon A. and Leandre R. Fabrigar. 1997. "Designing Rating Scales for Effective Measurement in Surveys." In *Survey Measurement and Process Quality*, edited by Lars Lyberg, Paul Biemer, Martin Collins, Edith de Leeuw, Cathryn Dippo, Norbert Schwarz, and Dennis Trewin, pp. 141-164. New York: John Wiley & Sons.
- La Porta, Rafael, Florencio Lopez-de-Silanes, Andrei Shleifer, and Robert W. Vishny. 1997. "Trust in Large Organizations." *American Economic Review* 87: 333-338.
- Lei, P-W. 2009. "Evaluating Estimation Methods for Ordinal Data in Structural Equation Modeling." *Quality and Quantity* 43: 495-507.
- Liu, Yu, Roger E. Millsap, Stephen G. West, Jenn-Yun Tein, Rika Tanaka, and Kevin J. Grimm. 2017. "Testing Measurement Invariance in Longitudinal Data with Ordered-Categorical Measures." *Psychological Methods* 22: 486-506.
- Lo Iacono, Sergio. 2018. "Does Community Social Embeddedness Promote Generalized Trust? An Experimental Test of the Spillover Effect." *Social Science Research* 73: 126-145.
- Lubke, Gitta H. and Bengt O. Muthén. 2004. "Applying Multigroup Confirmatory Factor Models for Continuous Outcomes to Likert Scale Data Complicates Meaningful Group Comparisons." *Structural Equation Modeling* 11: 514-534.

- Lundmark, Sebastian, Mikael Gilljam, and Stefan Dahlberg. 2016. "Measuring Generalized Trust: An Examination of Question Wording and the Number of Scale Points." *Public Opinion Quarterly* 80: 26-43.
- Ma, Debbie S., Joshua Correll, and Bernd Wittenbrink. 2015. "The Chicago Face Database: A Free Stimulus Set of Faces and Norming Data." *Behavior Research Methods* 47: 1122-1135.
- Messner, Steven F, Richard Rosenfeld, and Eric P. Baumer. 2004. "Dimensions of Social Capital and Rates of Criminal Homicide." *American Sociological Review* 69: 882-903.
- Mewes, Jan, Malcolm Fairbrother, Giuseppe Nicola Giordano, Cary Wu, and Rima Wilkes. 2021. "Experiences Matter: A Longitudinal Study of Individual-Level Sources of Declining Social Trust in the United States." *Social Science Research* 95: 102537.
- Miller, Alan S. and Tomoko Mitamura. 2003. "Are Surveys on Trust Trustworthy?" *Social Psychology Quarterly* 66: 60-70.
- Muthén, Bengt O. and David Kaplan. 1985. "A Comparison of Some Methodologies for the Factor Analysis of Non-Normal Likert Variables." *British Journal of Mathematical and Statistical Psychology* 38: 171-189.
- Muthén, Bengt O. and Linda K. Muthén. 2017. *Mplus 8* [Computer software]. Los Angeles: Authors.
- Naef, Michael, and Jürgen Schupp. 2009. "Measuring Trust: Experiments and Surveys in Contrast and Combination." *IZA DP No. 4087*.
- Nannestad, Peter. 2008. "What Have we Learned about Generalized Trust, if Anything?" *Annual Review of Political Science* 11: 413-436.
- Newton, Kenneth and Sonja Zmerli. 2011. "Three Forms of Trust and Their Association." *European Political Science Review* 3: 169-200.
- Nunnally, Jum. 1978. *Psychometric Theory*. New York: McGraw-Hill.
- Paxton, Pamela. 1999. "Is Social Capital Declining in the United States? A multiple indicator assessment." *American Journal of Sociology* 105: 88-127.
- Paxton, Pamela. 2002. "Social Capital and Democracy: An Interdependent Relationship." *American Sociological Review* 67: 254-277.
- Paxton, Pamela. 2007. "Association Memberships and Generalized Trust: A Multilevel Model across 31 countries." *Social Forces* 86: 47-76.
- Paxton, Pamela, and Jennifer Glanville, 2015. "Is Trust Rigid or Malleable? A Laboratory Experiment." *Social Psychology Quarterly* 78: 194-204.
- Peysakhovich, Alexander, Martin A. Nowak, and David G. Rand. 2014. "Humans Display a 'Cooperative Phenotype' that is Domain General and Temporally Stable." *Nature Communications* 5: 4939.
- Putnam, Robert. 1993. *Making Democracy Work: Civic Tradition in Modern Italy*. Princeton University Press, Princeton, NJ.
- Putnam, Robert. 2000. *Bowling Alone: The Collapse and Revival of American Community*. Free Press, New York, NY.
- Putnick, Diane L. and Marc H. Bornstein. 2016. "Measurement Invariance Conventions and Reporting: The State of the Art and Future Directions for Psychological Research." *Developmental Review* 41: 71-90.
- Reeskens, Tim and Marc Hooghe. 2008. "Cross-Cultural Measurement Equivalence of Generalized Trust. Evidence from the European Social Survey (2002 and 2004)." *Social Indicators Research* 85: 515-532..

- Robinson, Robert V. and Elton F. Jackson. 2001. "Is Trust Declining in America? An Age-Period-Cohort Analysis." *Social Science Research* 30: 117-145.
- Robbins, Blaine. 2011. "Neither Government nor Community Alone: A Test of State-Centered Models of Generalized Trust." *Rationality and Society* 23: 304-346.
- Robbins, Blaine. 2012a. "Institutional Quality and Generalized Trust: A Nonrecursive Causal Model." *Social Indicators Research* 107: 234-258.
- Robbins, Blaine. 2012b. "A Blessing and a Curse? Political Institutions in the Growth and Decay of Generalized Trust: A Cross-National Panel Analysis, 1980 to 2009." *PLoS ONE* 7: e35120.
- Robbins, Blaine. 2016a. "From the General to the Specific: How Social Trust Motivates Relational Trust." *Social Science Research* 55: 16-30.
- Robbins, Blaine. 2016b. "What is Trust? A Multidisciplinary Review, Critique, and Synthesis." *Sociology Compass* 10: 972-986.
- Robbins, Blaine G. 2019. "Measuring Generalized Trust: Two New Approaches." *Sociological Methods & Research*.
- Rosenberg, M. 1956. "Misanthropy and Political Ideology." *American Sociological Review* 21: 690-695.
- Rosenfeld, Richard, Steven F. Messner, and Eric P. Baumer. 2001. "Social Capital and Homicide." *Social Forces* 80: 283-309.
- Rothstein, Bo. 2011. *The Quality of Government: Corruption, Social Trust, and Inequality in International Perspective*. Chicago, IL: University of Chicago Press.
- Rothstein, Bo and Eric Uslaner. 2005. "All for All: Equality, Corruption, and Social Trust." *World Politics* 58 41-72."
- Rotter, J. B., 1967. "A New Scale for the Measurement of Interpersonal Trust." *Journal of Personality* 35: 651-665.
- Sapienza, Paola, Anna Toldra-Simats, and Luigi Zingales. 2013. "Understanding Trust." *Economic Journal* 123: 1313-1332.
- Sønderskov, Kim and Peter Dinesen. 2016. "Trusting the State, Trust Each Other? The Effect of Institutional Trust on Social Trust." *Political Behavior* 38: 179-202.
- Sturgis, Patrick and Patten Smith. 2010. "Assessing the Validity of Generalized Trust Questions: What Kind of Trust are we Measuring?" *International Journal of Public Opinion Research* 22: 74-92.
- Schwadel, Philip and Michael Stout. 2012. "Age, Period and Cohort Effects on Social Capital." *Social Forces* 91: 233-252.
- Stolle, Dietlind. 1998. "Bowling Together, Bowling Alone: The Development of Generalized Trust in Voluntary Organizations." *Political Psychology* 19: 497-525.
- Sztompka, Piotr. 1999. *Trust: A Sociological Theory*. Cambridge, UK: Cambridge University Press.
- Tabellini, Guido, 2010. "Culture and Institutions: Economic Development in the Regions of Europe." *Journal of the European Economic Association* 8: 677-716.
- Torpe, Lars and Henrik Lolle. 2011. "Identifying Social Trust in Cross-Country Analysis: Do we Really Measure the Same?" *Social Indicators Research* 103: 481-500.
- Traunmüller, Richard. 2011. "Moral Communities? Religion as a Source of Social Trust in a Multilevel Analysis of 97 German Regions." *European Sociological Review* 27: 346-363.
- Tsai, Ming-Chang, Leslie Laczkó, and Christian Bjørnskov. 2011. "Social Diversity, Institutions and Trust: A Cross-National Analysis." *Social Indicators Research* 101: 305-322.

- Uslaner, Eric. 2002. *The Moral Foundations of Trust*. Cambridge: Cambridge University Press.
- Uslaner, Eric. 2008a. *Corruption, Inequality, and the Rule of Law*. Cambridge, UK: Cambridge University Press.
- Uslaner, Eric. 2008b. "Where You Stand Depends Upon Where Your Grandparents Sat: The Inheritability of Generalized Trust." *Public Opinion Quarterly* 72: 725-740.
- Uslaner, Eric. 2011. "Generalized Trust Questions." In John H. Aldrich and Kathleen M. McGraw, eds., *Improving Public Opinion Surveys: Interdisciplinary Innovation and the American National Election Studies*, pp. 101-114. Princeton, NJ: Princeton University Press.
- Uslaner, Eric M. 2015. "Measuring Generalized Trust: In Defense of the 'Standard' Question." In Fergus Lyon, Guido Möllering, and Mark N.K. Saunders, eds., *Handbook of Research Methods on Trust*, 2nd Edition, pp. 97-106. Cheltenham, UK: Elgar.
- Van Hoorn, Andre. 2014. "Trust Radius versus Trust Level: Radius of Trust as a Distinct Trust Construct." *American Sociological Review* 79: 1256-1259.
- Vandenberg, Robert J. 2002. "Toward a Further Understanding of and Improvement in Measurement Invariance Methods and Procedures." *Organizational Research Methods* 5: 139-158.
- Vandenberg, Robert J. and Charles Lance. 2000. "A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research." *Organizational Research Methods* 3: 4-70.
- Welch, Michael, David Sikkink, Eric Sartain, and Carolyn Bond. 2004. "Trust in God and Trust in Man: The Ambivalent Role of Religion in Shaping Dimensions of Social Trust." *Journal for the Scientific Study of Religion* 43: 317-343.
- Wilkes, Rima. 2011. "Re-thinking the Decline in Trust: A Comparison of Black and White Americans." *Social Science Research* 40 1596-1610.
- Wu, Cary. 2020. "Does Migration Affect Trust? Internal Migration and the Stability of Trust Among Americans." *Sociological Quarterly* 61: 523-543.
- You, Jong-sung. 2012. "Social Trust: Fairness Matters More than Homogeneity." *Political Psychology* 33: 701-721.
- Zak, Paul J. and Stephen Knack. 2001. "Trust and Growth." *The Economic Journal* 111: 295-321.
- Zmerli, Sonja and Kenneth Newton. 2008. "Social Trust and Attitudes Toward Democracy." *Public Opinion Quarterly* 72: 706-724.

Table 1. Descriptive Statistics of SFT, IST, MST, and GST in Studies 1, 2, and 3

Factors and Indicators	Study 1				Study 2				Study 3			
	α	M	SD	N	α	M	SD	N	α	M	SD	N
<i>Stranger-face trust (SFT)</i>	.962				.968				.954			
White female (WF)	.922	0.78	0.82	1813	.926	0.59	0.72	1519	.779	0.86	0.63	139
White male (WM)	.928	0.57	0.73	1819	.923	0.43	0.64	1530	.840	0.62	0.58	138
Black female (BF)	.926	0.77	0.83	1799	.924	0.58	0.73	1514	.781	0.84	0.61	138
Black male (BM)	.929	0.80	0.85	1810	.931	0.61	0.76	1507	.784	0.79	0.57	137
Latin female (LF)	.924	0.75	0.82	1808	.919	0.56	0.70	1519	.765	0.87	0.63	138
Latin male (LM)	.925	0.73	0.81	1805	.930	0.56	0.72	1510	.795	0.73	0.55	138
<i>Imaginary stranger trust (IST)</i>	.929				.921				.765			
Secret (SECRET)		0.59	0.90	1870		0.39	0.72	1598		0.49	0.74	143
Loan (LOAN)		0.54	0.88	1881		0.37	0.71	1595		0.41	0.70	142
Child (CHILD)		0.51	0.90	1896		0.33	0.68	1618		0.41	0.62	142
Advice (ADVICE)		0.77	0.92	1848		0.60	0.79	1563		1.06	0.79	141
<i>Misanthropy scale (MST)</i>	.686				.678				.675			
Trust (TRUST)		0.33	0.47	1990		0.31	0.46	1687		0.37	0.48	136
Fair (FAIR)		0.54	0.88	1881		0.73	0.44	1603		0.58	0.49	125
Helpful (HELP)		0.51	0.90	1896		0.52	0.49	1622		0.35	0.48	137
<i>Generalized social trust scale (GST)</i>	.741				.787				.783			
Meet first time (FIRST)		1.24	0.82	1859		1.14	0.77	1585		1.09	0.69	137
Another religion (RELI)		1.89	0.72	1756		1.86	0.67	1467		1.96	0.66	119
Another nationality (NATION)		1.86	0.70	1768		1.81	0.69	1490		1.87	0.63	121

Table 2. Confirmatory Factor Analysis Assessing Convergent and Discriminant Validity in Studies 1, 2, and 3

Latent Factor	Indicator	Study 1			Study 2			Study 3		
		Loading	VE	AVE	Loading	VE	AVE	Loading	VE	AVE
SFT				.806			.846			.786
	WF	0.890***	.792		0.917***	.842		0.866***	.750	
	WM	0.855***	.731		0.870***	.756		0.823***	.678	
	BF	0.901***	.811		0.927***	.859		0.879***	.773	
	BM	0.903***	.815		0.934***	.872		0.923***	.852	
	LF	0.922***	.851		0.929***	.862		0.900***	.809	
	LM	0.916***	.840		0.943***	.890		0.925***	.856	
IST				.763			.752			.469
	SECRET	0.895***	.800		0.876***	.767		0.609***	.371	
	LOAN	0.906***	.821		0.917***	.841		0.794***	.630	
	CHILD	0.880***	.775		0.895***	.801		0.749***	.561	
	ADVICE	0.811***	.657		0.775***	.601		0.563***	.317	
MST				.421			.413			.390
	TRUST	0.658***	.433		0.629***	.395		0.765***	.585	
	FAIR	0.558***	.311		0.580***	.336		0.541***	.293	
	HELP	0.722***	.521		0.713***	.508		0.541***	.292	
GST				.534			.622			.614
	FIRST	0.641***	.411		0.599***	.359		0.542***	.294	
	RELI	0.775***	.601		0.869***	.755		0.886***	.785	
	NATION	0.768***	.590		0.867***	.752		0.875***	.765	
		<i>r</i>	<i>r</i> ²		<i>r</i>	<i>r</i> ²		<i>r</i>	<i>r</i> ²	
	Corr (SFT, IST)	.714	.509		.721	.519		.614	.376	
	Corr (SFT, MST)	.312	.097		.216	.046		.365	.133	
	Corr (SFT, GST)	.427	.182		.343	.117		.255	.065	
	Corr (IST, MST)	.254	.064		.222	.049		.368	.135	
	Corr (IST, GST)	.334	.111		.291	.084		.205	.042	
	Corr (MST, GST)	.600	.360		.609	.370		.523	.273	
		<i>N</i> = 2030			<i>N</i> = 1745			<i>N</i> = 146		
		χ^2 (98) = 651.310, <i>p</i> < .001			χ^2 (98) = 445.427, <i>p</i> < .001			χ^2 (98) = 185.776, <i>p</i> < .001		
		RMSEA = .053			RMSEA = .045			RMSEA = .078		
		90% C.I. RMSEA = [.049, .057]			90% C.I. RMSEA = [.041, .049]			90% C.I. RMSEA = [.061, .095]		
		CFI = .960			CFI = .969			CFI = .921		
		SRMR = .070			SRMR = .068			SRMR = .075		

*** *p* < .001

Notes : standardized factor loadings. Indicators for SFT, IST, MST, and GST treated as continuous (factor variances, item intercepts, and residual variances freely estimated but not shown). VE = variance extracted (i.e., the R^2 or the Squared Multiple Correlation); AVE = average variance extracted (the average of the VEs for each latent construct). All correlations between latent constructs were statistically significant at the *p* < .05 level.

Table 3. Tests of Measurement Invariance of SFT by Gender, Race, Education, and Age-Groups in Study 1

	<i>df</i>	χ^2	SRMR	RMSEA	CFI	Δdf	$\Delta\chi^2$	Δ SRMR	Δ RMSEA	Δ CFI
<i>Gender (N = 1874)</i>										
1. Configural Invariance	18	52.555*	.012	.045	.990					
1 versus 2						5	14.282*	.008	.000	-.003
2. Metric Invariance	23	67.010*	.020	.045	.987					
2 versus 3						5	12.484*	.000	.000	-.002
3. Scalar Invariance	28	80.671*	.020	.045	.985					
<i>Race (N = 1877)</i>										
1. Configural Invariance	18	54.739*	.012	.047	.990					
1 versus 2						5	5.293	.003	-.003	-.002
2. Metric Invariance	23	65.596*	.015	.044	.988					
2 versus 3						5	3.718	.001	-.002	-.001
3. Scalar Invariance	28	75.269*	.016	.042	.987					
<i>Education (N = 1865)</i>										
1. Configural Invariance	36	71.151*	.015	.046	.990					
1 versus 2						15	27.658*	.013	-.001	-.004
2. Metric Invariance	51	99.832*	.028	.045	.986					
2 versus 3						15	23.777	.000	-.001	-.003
3. Scalar Invariance	66	126.550*	.028	.044	.983					
<i>Age (N = 1877)</i>										
1. Configural Invariance	36	68.924*	.014	.044	.990					
1 versus 2						15	13.112	.012	-.003	-.002
2. Metric Invariance	51	90.263*	.022	.041	.988					
2 versus 3						15	13.458	.001	-.003	-.002
3. Scalar Invariance	66	111.035*	.023	.038	.986					

Note : SRMR = standardized root mean squared residual; RMSEA = root mean squared error of approximation; CFI = comparative fit index; $\Delta\chi^2$ = Satorra-Bentler scaled χ^2 difference test; * $p < .05$

Table 4. Tests of Measurement Invariance of IST by Gender, Race, Education, and Age-Groups in Study 1

	<i>df</i>	χ^2	SRMR	RMSEA	CFI	Δdf	$\Delta\chi^2$	Δ SRMR	Δ RMSEA	Δ CFI
<i>Gender (N = 1936)</i>										
1. Configural Invariance	4	6.741	.007	.027	.999					
1 versus 2						3	4.339	.008	-.002	-.001
2. Metric Invariance	7	11.299	.015	.025	.998					
2 versus 3						3	10.667*	.003	.007	-.003
3. Scalar Invariance	10	19.659*	.018	.032	.995					
<i>Race (N = 1939)</i>										
1. Configural Invariance	4	19.661*	.011	.064	.993					
1 versus 2						3	10.444*	.012	-.004	-.004
2. Metric Invariance	7	31.462*	.023	.060	.989					
2 versus 3						3	4.636	.001	-.005	-.002
3. Scalar Invariance	10	39.088*	.024	.055	.987					
<i>Education (N = 1926)</i>										
1. Configural Invariance	8	17.902*	.012	.051	.996					
1 versus 2						9	7.923	.009	-.013	-.001
2. Metric Invariance	17	28.963*	.021	.038	.995					
2 versus 3						9	17.061*	.003	.001	-.004
3. Scalar Invariance	26	45.525*	.024	.039	.991					
<i>Age (N = 1939)</i>										
1. Configural Invariance	8	12.981	.001	.036	.998					
1 versus 2						9	26.241*	.036	.015	-.007
2. Metric Invariance	17	36.419*	.037	.049	.991					
2 versus 3						9	31.168*	.009	.006	-.009
3. Scalar Invariance	26	63.572*	.046	.055	.982					

Note : SRMR = standardized root mean squared residual; RMSEA = root mean squared error of approximation; CFI = comparative fit index; $\Delta\chi^2$ = Satorra-Bentler scaled χ^2 difference test; * $p < .05$

Table 5. Tests of Measurement Invariance of MST by Gender, Race, Education, and Age-Groups in Study 1

	<i>df</i>	χ^2	SRMR	RMSEA	CFI	Δdf	$\Delta\chi^2$	Δ SRMR	Δ RMSEA	Δ CFI
<i>Gender (N = 2007)</i>										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						2	18.374*	.033	.090	-.023
2. Metric Invariance	2	18.374*	.033	.090	.977					
2 versus 3						2	29.351*	.007	.015	-.038
3. Scalar Invariance	4	48.065*	.040	.105	.939					
<i>Race (N = 2010)</i>										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						2	1.506	.009	.000	.000
2. Metric Invariance	2	1.506	.009	.000	1.000					
2 versus 3						2	12.739*	.023	.049	-.014
3. Scalar Invariance	4	13.841*	.023	.049	.986					
<i>Education (N = 1996)</i>										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						6	30.650*	.044	.091	-.034
2. Metric Invariance	6	30.650*	.044	.091	.966					
2 versus 3						6	30.069*	.006	-.001	-.033
3. Scalar Invariance	12	60.694*	.050	.090	.933					
<i>Age (N = 2010)</i>										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						6	14.805*	.031	.054	-.013
2. Metric Invariance	6	14.805*	.031	.054	.987					
2 versus 3						6	46.578*	.019	.037	-.059
3. Scalar Invariance	12	62.208	.050	.091	.928					

Note : SRMR = standardized root mean squared residual; RMSEA = root mean squared error of approximation; CFI = comparative fit index; $\Delta\chi^2$ = Satorra-Bentler scaled χ^2 difference test; * $p < .05$

Table 6. Tests of Measurement Invariance of GST by Gender, Race, Education, and Age-Groups in Study 1

	<i>df</i>	χ^2	SRMR	RMSEA	CFI	Δdf	$\Delta \chi^2$	Δ SRMR	Δ RMSEA	Δ CFI
<i>Gender (N = 1954)</i>										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						2	3.763	.021	.030	-.002
2. Metric Invariance	2	3.763	.021	.030	.998					
2 versus 3						2	53.514*	.023	.081	-.065
3. Scalar Invariance	4	52.095	.044	.111	.933					
<i>Race (N = 1957)</i>										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						2	3.036	.016	.023	-.001
2. Metric Invariance	2	3.036	.016	.023	.999					
2 versus 3						2	0.894	.002	-.016	.001
3. Scalar Invariance	4	4.194	.018	.007	1.000					
<i>Education (N = 1944)</i>										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						6	6.323	.022	.011	.000
2. Metric Invariance	6	6.323	.022	.011	1.000					
2 versus 3						6	32.390*	.024	.053	-.049
3. Scalar Invariance	12	36.025*	.046	.064	.967					
<i>Age (N = 1957)</i>										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						6	3.505	.019	.000	.000
2. Metric Invariance	6	3.505	.019	.000	1.000					
2 versus 3						6	19.496*	.020	.039	-.013
3. Scalar Invariance	12	21.010	.039	.039	.987					

Note : SRMR = standardized root mean squared residual; RMSEA = root mean squared error of approximation; CFI = comparative fit index; $\Delta \chi^2$ = Satorra-Bentler scaled χ^2 difference test; * $p < .05$

Table 7. Reliability, Validity, and Measurement Invariance of SFT, IST, MST, and GST

	SFT	IST	MST	GST	Inequalities
<i>Reliability</i>					
Internal consistency	Good	Good	Poor-to-Acceptable	Acceptable	SFT = IST > GST > MST
Test-retest reliability	Poor	Poor	Good	Poor-to-Acceptable	MST > GST > SFT > IST
Longitudinal MI/E	Acceptable	Acceptable	Acceptable	Acceptable	SFT = IST = MST = GST
<i>Validity</i>					
Convergent validity	Acceptable-to-Good	Acceptable-to-Good	Poor-to-Acceptable	Poor-to-Acceptable	SFT = IST > MST = GST
Discriminant validity	Acceptable	Poor-to-Acceptable	Poor-to-Acceptable	Poor-to-Acceptable	SFT > IST = MST = GST
Concurrent validity ^A	Acceptable	Acceptable	Acceptable	Acceptable	SFT = IST = MST = GST
Predictive validity	Acceptable	Poor	Acceptable	Poor	SFT = MST > IST = GST
Known-groups validity ^B	Acceptable	Acceptable	Acceptable	Acceptable	SFT = IST = MST = GST
<i>Measurement Invariance/Equivalence</i>					
Multiple group (gender)	Acceptable	Acceptable	Poor-to-Acceptable	Poor-to-Acceptable	SFT = IST > MST = GST
Multiple group (race)	Acceptable	Acceptable	Poor-to-Acceptable	Acceptable	SFT = IST = GST > MST
Multiple group (education)	Acceptable	Acceptable	Poor-to-Acceptable	Poor-to-Acceptable	SFT = IST > MST = GST
Multiple group (age-groups)	Acceptable	Acceptable	Poor-to-Acceptable	Poor-to-Acceptable	SFT = IST > MST = GST
Multiple group (samples)	Acceptable	Acceptable	Acceptable	Acceptable	SFT = IST = MST = GST

^A Given the results of SFT and IST, tests of concurrent validity executed by Robbins (2019) suggest that the correlations between criterion variables and MST/GST are biased.

^B In Robbins (2019), known-groups validity was established for SFT, IST, MST, and GST, but some known differences failed to replicate (e.g., education) or were unexpectedly in the opposite direction (i.e., age).

Note : Good = values of a validation test close to perfect; Acceptable = values of a validation test greater than common cutoffs; Poor = values of a validation test less than common cutoffs; SFT = stranger face trust scale; IST = imaginary stranger trust scale; MST = misanthropy scale; GST = generalized social trust scale; MI/E = measurement invariance/equivalence.

Figure 1. Scatterplots of the Relation Between the Proportion of Tokens Sent in an Investment Game (y-axis) and Latent Factor Scores of SFT (Panel A), IST (Panel B), MST (Panel C), and GST (Panel D)

Note: $N = 45$. Best fit linear regression lines are solid red, 95% CIs are grey dashed lines, and individual observations are blue circles. Response propensity weighting used for all estimates.

Figure 2. Scatterplots of the Relation Between the Proportion of Tokens Returned in an Investment Game (y-axis) and Latent Factor Scores of SFT (Panel A), IST (Panel B), MST (Panel C), and GST (Panel D)

Note: $N = 35$. Best fit linear regression lines are solid red, 95% CIs are grey dashed lines, and individual observations are blue circles. Response propensity weighting used for all estimates.

Supplemental Material

An Empirical Comparison of Four Generalized Trust Scales: Test-Retest Reliability, Measurement Invariance, Predictive Validity, and Replicability

Contents

1. Stranger Face Trust scale	
1.1. Figure S1. Six human faces.....	2
1.2. Figure S2. Example SFT instrument.....	3
2. Study 1 – Robustness	
2.1. Table S1. Four-factor CFA with categorical indicators	4
2.2. Table S2. Higher-order CFA of SFT with categorical indicators	5
2.3. Table S3. MI/E of higher-order SFT by sociodemographic groups.....	6
2.4. Table S4. MI/E of IST with categorical indicators	7
2.5. Table S5. MI/E of MST with categorical indicators	8
2.6. Table S6. MI/E of GST with categorical indicators.....	9
3. Study 2 – Sample Characteristics, Test-Retest Reliability, and MI/E	
3.1. Table S7. Comparison of study 1 and study 2 sample characteristics	10
3.2. Table S8. Test-retest reliability of SFT, IST, MST, and GST	11
3.3. Table S9. MI/E of SFT by sociodemographic groups.....	12
3.4. Table S10. MI/E of IST by sociodemographic groups	13
3.5. Table S11. MI/E of MST by sociodemographic groups	14
3.6. Table S12. MI/E of GST by sociodemographic groups	15
3.7. Table S13. Longitudinal MI/E of SFT, IST, MST, and GST.....	16
4. Study 2 – Robustness	
4.1. Table S14. MI/E of higher-order SFT by sociodemographic groups.....	17
4.2. Table S15. MI/E of IST with categorical indicators	18
4.3. Table S16. MI/E of MST with categorical indicators	19
4.4. Table S17. MI/E of GST with categorical indicators.....	20
4.5. Table S18. Longitudinal MI/E of higher-order SFT	21
4.6. Table S19. Longitudinal MI/E of IST, MST, and GST with categorical indicators	22
4.7. Table S20. Longitudinal MI/E of SFT, IST, MST, and GST with listwise deletion.....	23
5. Study 3 – MI/E	
5.1. Table S21. MI/E of SFT, IST, MST, and GST across samples	24
6. Study 3 – Robustness	
6.1. Table S22. MI/E of IST, MST, and GST across samples with categorical indicators	25
6.2. Table S23. Robustness of predictive validity, OLS regression.....	26
6.3. Table S24. Robustness of predictive validity, fractional regression	27
7. References	
7.1. References.....	28

1. Stranger Face Trust Scale

Figure S1. Stranger Faces used for the SFT Scale

Notes. The six faces were drawn from the Chicago Face Database (Ma et al. 2014). The faces used in the present study were WF-228 (Panel A), WM-225 (Panel B), BF-214 (Panel C), BM-021 (Panel D), LF-240 (Panel E), and LM-227 (Panel F).

Please identify how much you would trust this stranger for each of the following:

Please select one answer per row.

	Trust completely	Trust somewhat	Do not trust very much	Do not trust at all	Don't know
To repay a loan of one thousand dollars?	☐	☐	☐	☐	☐
To keep a secret that is damaging to your reputation?	☐	☐	☐	☐	☐
To provide advice about how best to manage your money?	☐	☐	☐	☐	☐
To look after a child, family member, or loved one while you are away?	☐	☐	☐	☐	☐

>>

Figure S2. Example of SFT Instrument in Qualtrics

Notes. The figure includes a human face (BM-021), the four trust domains, and the rating scales

2. Study 1 – Robustness

Table S1. Confirmatory Factor Analysis Assessing Convergent and Discriminant Validity of SFT, IST, MST, and GST in Studies 1, 2, and 3

Latent Factor	Indicator	Study 1 - MLR			Study 2 - MLR			Study 2 - WLSMV			Study 3 - MLR		
		Loading	VE	AVE	Loading	VE	AVE	Loading	VE	AVE	Loading	VE	AVE
SFT				.807			.846			.845			.787
	WF	0.890***	.793		0.917***	.841		0.920***	.846		0.866***	.750	
	WM	0.854***	.729		0.868***	.754		0.882***	.777		0.823***	.678	
	BF	0.901***	.813		0.927***	.859		0.928***	.862		0.880***	.774	
	BM	0.903***	.816		0.935***	.874		0.913***	.834		0.923***	.853	
	LF	0.923***	.852		0.928***	.861		0.938***	.879		0.900***	.810	
	LM	0.916***	.840		0.944***	.890		0.935***	.875		0.926***	.857	
IST				.849			.835			.867			.593
	SECRET	0.931***	.866		0.900***	.810		0.935***	.875		0.690***	.476	
	LOAN	0.950***	.902		0.957***	.917		0.964***	.930		0.877***	.769	
	CHILD	0.930***	.865		0.945***	.892		0.950***	.903		0.844***	.712	
	ADVICE	0.875***	.765		0.849***	.721		0.874***	.763		0.645***	.417	
MST				.613			.645			.669			.507
	TRUST	0.846***	.716		0.859***	.738		0.959***	.919		0.730***	.533	
	FAIR	0.664***	.441		0.738***	.545		0.665***	.443		0.651***	.424	
	HELP	0.827***	.683		0.808***	.652		0.804***	.647		0.751***	.564	
GST				.642			.716			.778			.738
	FIRST	0.713***	.508		0.693***	.480		0.901***	.811		0.622***	.387	
	RELI	0.852***	.726		0.914***	.835		0.877***	.768		0.915***	.837	
	NATION	0.833***	.694		0.913***	.834		0.870***	.756		0.995***	.990	
		<i>r</i>	<i>r</i> ²		<i>r</i>	<i>r</i> ²		<i>r</i>	<i>r</i> ²		<i>r</i>	<i>r</i> ²	
	Corr(SFT, IST)	.717	.514		.722	.521		.718	.515		.653	.426	
	Corr(SFT, MST)	.325	.105		.247	.061		.241	.058		.384	.147	
	Corr(SFT, GST)	.458	.209		.393	.154		.402	.161		.270	.072	
	Corr(IST, MST)	.275	.075		.307	.094		.290	.084		.323	.104	
	Corr(IST, GST)	.385	.148		.381	.145		.457	.208		.273	.074	
	Corr(MST, GST)	.658	.432		.691	.477		.660	.435		.624	.389	
		<i>N</i> = 2030			<i>N</i> = 1745			<i>N</i> = 1745			<i>N</i> = 146		
		AIC = 40561			AIC = 27583			χ^2 (98) = 1235.650, <i>p</i> < .001			AIC = 2682		
		BIC = 40887			BIC = 27900			RMSEA = .082			BIC = 2855		
								90% C.I. RMSEA = [.078, .086]					
								CFI = .942					
								SRMR = .062					

*** *p* < .001

Notes : standardized factor loadings. WLSMV failed to converge for Study 1 and Study 3. Indicators for SFT treated as continuous, all other indicators treated as categorical (factor variances, item intercepts, residual variances, and item thresholds freely estimated but not shown). MLR = robust maximum likelihood with numerical integration; WLSMV = mean-and-variance adjusted weighted least squares; VE = variance extracted (i.e., the *R*² or the Squared Multiple Correlation); AVE = average variance extracted (the average of the VEs for each latent construct). All correlations between latent constructs were statistically significant at the *p* < .05 level.

Table S2. Higher-Order Confirmatory Factor Analysis of SFT in Studies 1 and 2

Latent Factor	Indicator	Study 1		Study 2	
		MLR (Continuous)	WLSMV (Categorical)	MLR (Continuous)	WLSMV (Categorical)
		Loading	Loading	Loading	Loading
White Female (WF)					
	WF SECRET	0.885***	0.930***	0.879***	0.928***
	WF LOAN	0.898***	0.943***	0.919***	0.958***
	WF CHILD	0.853***	0.905***	0.879***	0.934***
	WF ADVICE	0.814***	0.893***	0.804***	0.905***
White Male (WM)					
	WM SECRET	0.873***	0.928***	0.888***	0.943***
	WM LOAN	0.896***	0.948***	0.903***	0.951***
	WM CHILD	0.880***	0.942***	0.878***	0.935***
	WM ADVICE	0.829***	0.906***	0.815***	0.907***
Black Female (BF)					
	BF SECRET	0.877***	0.922***	0.884***	0.932***
	BF LOAN	0.913***	0.948***	0.908***	0.952***
	BF CHILD	0.874***	0.926***	0.881***	0.931***
	BF ADVICE	0.815***	0.898***	0.815***	0.906***
Black Male (BM)					
	BM SECRET	0.874***	0.922***	0.881***	0.932***
	BM LOAN	0.909***	0.948***	0.916***	0.956***
	BM CHILD	0.885***	0.935***	0.893***	0.941***
	BM ADVICE	0.818***	0.893***	0.836***	0.927***
Latin Female (LF)					
	LF SECRET	0.883***	0.930***	0.884***	0.932***
	LF LOAN	0.898***	0.939***	0.907***	0.949***
	LF CHILD	0.854***	0.907***	0.860***	0.922***
	LF ADVICE	0.832***	0.908***	0.819***	0.907***
Latin Male (LM)					
	LM SECRET	0.877***	0.925***	0.874***	0.928***
	LM LOAN	0.903***	0.941***	0.917***	0.958***
	LM CHILD	0.868***	0.928***	0.891***	0.943***
	LM ADVICE	0.819***	0.904***	0.836***	0.922***
SFT					
	WF	0.933***	0.958***	0.957***	0.977***
	WM	0.886***	0.930***	0.900***	0.946***
	BF	0.937***	0.962***	0.962***	0.984***
	BM	0.946***	0.968***	0.973***	0.992***
	LF	0.964***	0.983***	0.965***	0.986***
	LM	0.955***	0.976***	0.981***	0.996***
		$N = 1877$	$N = 1877$	$N = 1563$	$N = 1563$
		$\chi^2 (246) = 3399.112$	$\chi^2 (246) = 4805.828$	$\chi^2 (246) = 3165.924$	$\chi^2 (246) = 4300.666$
		$p < .001$	$p < .001$	$p < .001$	$p < .001$
		RMSEA = .083	RMSEA = .099	RMSEA = .087	RMSEA = .103
		CFI = .844	CFI = .977	CFI = .813	CFI = .980
		SRMR = .043	SRMR = .036	SRMR = .047	SRMR = .039

*** $p < .001$

Notes : standardized factor loadings. MLR and WLSMV failed to converge for Study 3. Higher-order CFAs estimated using robust maximum likelihood (MLR) for continuous indicators and mean-and-variance adjusted weighted least squares (WLSMV) for categorical indicators.

Table S3. Tests of Measurement Invariance of Higher-Order SFT by Gender, Race, Education, and Age-Groups in Study 1

	<i>df</i>	χ^2	SRMR	RMSEA	CFI	Δdf	$\Delta\chi^2$	Δ SRMR	Δ RMSEA	Δ CFI
<i>Gender (N = 1874)</i>										
1. Configural	498	4258.871*	.192	.090	.815					
1 versus 2						18	68.814*	.005	-.001	-.004
2. First-Order Metric	516	4371.384*	.197	.089	.811					
2 versus 3						5	18.574*	.000	.000	-.001
3. First-and-Second-Order Metric	521	4400.241*	.197	.089	.810					
3 versus 4						18	109.170*	.001	.000	-.006
4. First-Order Scalar	539	4533.412*	.198	.089	.804					
4 versus 5						5	9.725	.000	.000	-.001
5. First-and-Second-Order Scalar	544	4561.683*	.198	.089	.803					
<i>Race (N = 1877)</i>										
1. Configural	498	4257.916*	.132	.090	.831					
1 versus 2						18	38.730*	.002	-.001	-.003
2. First-Order Metric	516	4357.954*	.134	.089	.828					
2 versus 3						5	3.845	.001	.000	-.001
3. First-and-Second-Order Metric	521	4376.582*	.135	.089	.827					
3 versus 4						18	45.890*	.010	-.001	-.004
4. First-Order Scalar	539	4479.105	.145	.088	.823					
4 versus 5						5	14.544*	.001	.000	-.001
5. First-and-Second-Order Scalar	544	4512.444	.146	.088	.822					
<i>Education (N = 1865)</i>										
1. Configural	990	5125.645*	.173	.095	.817					
1 versus 2						54	102.743*	-.003	-.001	-.006
2. First-Order Metric	1044	5316.434*	.170	.094	.811					
2 versus 3						15	31.557*	.001	-.001	-.001
3. First-and-Second-Order Metric	1059	5365.852*	.171	.093	.810					
3 versus 4						54	110.111*	.004	.000	-.006
4. First-Order Scalar	1113	5558.127*	.175	.093	.804					
4 versus 5						15	19.628	.000	-.001	-.002
5. First-and-Second-Order Scalar	1128	5606.114*	.175	.092	.802					
<i>Age (N = 1877)</i>										
1. Configural	990	5275.963*	.190	.096	.809					
1 versus 2						54	74.840*	.005	-.001	-.005
2. First-Order Metric	1044	5453.473*	.195	.095	.804					
2 versus 3						15	9.468	.000	-.001	-.001
3. First-and-Second-Order Metric	1059	5487.313	.195	.094	.803					
3 versus 4						54	180.734*	.012	.000	-.008
4. First-Order Scalar	1113	5718.024	.207	.094	.795					
4 versus 5						15	15.517	.000	.000	-.002
5. First-and-Second-Order Scalar	1128	5766.891*	.207	.094	.793					

Note : model identification and nested model comparisons follow guidelines outlined by Rudnev et al. (2018). SRMR = standardized root mean squared residual; RMSEA = root mean squared error of approximation; CFI = comparative fit index; $\Delta\chi^2$ = Satorra-Bentler scaled χ^2 difference test; * $p < .05$

Table S4. Tests of Measurement Invariance of IST by Gender, Race, Education, and Age-Groups in Study 1

	<i>df</i>	χ^2	SRMR	RMSEA	CFI	Δdf	$\Delta\chi^2$	Δ SRMR	Δ RMSEA	Δ CFI
<i>Gender (N = 1936)</i>										
1. Configural Invariance	4	8.458	.004	.034	1.000					
1 versus 2						4	8.911	.003	.001	.000
2. Threshold Invariance	8	17.309*	.007	.035	1.000					
2 versus 3						3	5.919	.000	-.001	.000
3. Loading Invariance	11	23.113*	.007	.034	1.000					
<i>Race (N = 1939)</i>										
1. Configural Invariance	4	29.890*	.007	.082	.999					
1 versus 2						4	5.312	.001	-.023	.000
2. Threshold Invariance	8	35.433*	.008	.059	.999					
2 versus 3						3	0.133	.000	-.012	.000
3. Loading Invariance	11	34.318*	.008	.047	.999					
<i>Education (N = 1926)</i>										
1. Configural Invariance	8	21.341*	.007	.059	1.000					
1 versus 2						12	11.016	.002	-.026	.000
2. Threshold Invariance	20	30.735	.009	.033	1.000					
2 versus 3						9	9.914	.002	-.005	.000
3. Loading Invariance	29	40.314	.011	.028	1.000					
<i>Age (N = 1939)</i>										
1. Configural Invariance	8	21.384*	.006	.059	1.000					
1 versus 2						12	31.391*	.008	-.002	.001
2. Threshold Invariance	20	51.974*	.014	.057	.999					
2 versus 3						9	18.528*	.001	-.003	.000
3. Loading Invariance	29	69.791*	.015	.054	.999					

Note : model identification and nested model comparisons follow guidelines outlined by Wu and Estabrook (2016) and Svetina et al. (2020). Indicators treated as categorical and models estimated with mean-and-variance adjusted weighted least squares (WLSMV) using delta parameterization. SRMR = standardized root mean squared residual; RMSEA = root mean squared error of approximation; CFI = comparative fit index; $\Delta\chi^2$ = two-step χ^2 difference test for WLSMV; * $p < .05$

Table S5. Tests of Measurement Invariance of MST by Gender, Race, Education, and Age-Groups in Study 1

	<i>df</i>	χ^2	SRMR	RMSEA	CFI	Δdf	$\Delta\chi^2$	Δ SRMR	Δ RMSEA	Δ CFI
<i>Gender (N = 2007)</i>										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						2	9.570*	.024	.061	-.004
2. Threshold-Loading Invariance	2	9.570*	.024	.061	.996					
<i>Race (N = 2010)</i>										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						2	3.023	.016	.023	-.001
2. Threshold-Loading Invariance	2	3.024	.016	.023	.999					
<i>Education (N = 1996)</i>										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						6	26.834*	.045	.083	-.011
2. Threshold-Loading Invariance	6	26.835*	.045	.083	.989					
<i>Age (N = 2010)</i>										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						6	33.748*	.048	.096	-.015
2. Threshold-Loading Invariance	6	33.748*	.048	.096	.985					

Note : model identification and nested model comparisons follow guidelines outlined by Wu and Estabrook (2016). Indicators treated as categorical and models estimated with mean-and-variance adjusted weighted least squares (WLSMV) using delta parameterization.

SRMR = standardized root mean squared residual; RMSEA = root mean squared error of approximation; CFI = comparative fit index;

$\Delta\chi^2$ = two-step χ^2 difference test for WLSMV; * $p < .05$

Table S6. Tests of Measurement Invariance of GST by Gender, Race, Education, and Age-Groups in Study 1

	<i>df</i>	χ^2	SRMR	RMSEA	CFI	Δdf	$\Delta\chi^2$	Δ SRMR	Δ RMSEA	Δ CFI
<i>Gender (N = 1954)</i>										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						3	2.353	.022	.000	.000
2. Threshold Invariance	3	2.353	.004	.000	1.000					
2 versus 3						2	6.572*	.010	.031	-.001
3. Loading Invariance	5	9.659	.010	.031	.999					
<i>Race (N = 1957)</i>										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						3	5.017	.007	.026	-.001
2. Threshold Invariance	3	5.017	.007	.026	.999					
2 versus 3						2	2.957	.003	-.002	.000
3. Loading Invariance	5	7.825	.010	.024	.999					
<i>Education (N = 1944)</i>										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						9	5.681	.007	.000	.000
2. Threshold Invariance	9	5.681	.007	.000	1.000					
2 versus 3						6	9.231	.007	.011	.000
3. Loading Invariance	15	15.896	.014	.011	1.000					
<i>Age (N = 1957)</i>										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						9	55.062*	.024	.102	-.007
2. Threshold Invariance	9	55.063*	.024	.102	.993					
2 versus 3						6	14.393*	.003	-.020	-.001
3. Loading Invariance	15	64.220*	.027	.082	.992					

Note : model identification and nested model comparisons follow guidelines outlined by Wu and Estabrook (2016) and Svetina et al. (2020). Indicators treated as categorical and models estimated with mean-and-variance adjusted weighted least squares (WLSMV) using delta parameterization. SRMR = standardized root mean squared residual; RMSEA = root mean squared error of approximation; CFI = comparative fit index; $\Delta\chi^2$ = two-step χ^2 difference test for WLSMV; * $p < .05$

3. Study 2 – Sample Characteristics, Test-Retest Reliability, and MI/E

Table S7. Comparison of Study 1 and Study 2 Sample Characteristics

	Study 1	Study 2	Difference
Household Income			
Less than \$30,000	22.4	16	6.4
\$30,000 to \$74,999	35.2	36.1	.9
\$75,000 to \$124,999	24	27.5	3.5
\$125,000 Plus	18.3	20.4	2.1
Age			
18 - 34	29.9	25.9	4
35 - 49	24.9	24.2	.7
50 - 64	25	26.8	1.8
65 Plus	19.9	23.1	3.2
Race-Ethnicity			
Non-Hispanic White	81.3	83.4	2.3
Non-Hispanic Black	5.6	4.6	1
Hispanic	6.4	4.7	1.7
Non-Hispanic Asian and Pacific Islander	3.7	3.6	.1
Non-Hispanic Other race	3	3.6	.6
Education Status			
Less than High School	1.8	.7	1.1
High School Equivalent	36.7	25	11.7
Some College or Associate Degree	16.3	13.3	3
Bachelor's Degree	26.2	35.2	9
Graduate Degree	19	25.8	6.8
Marital Status			
Currently Married	54.2	56.5	2.3
Separated, Divorced, Widowed, or Single	45.8	43.5	2.3
Sex			
Male	49.5	48.4	1.1
Female	50.4	51.6	1.2
Average Difference			3

Table S8. Test-Retest Reliability of SFT, IST, MST, and GST in Study 2

	Panel Attrition - Missing Data Estimation		Panel Attrition - Listwise Deletion	
	MLR (Continuous)	WLSMV (Categorical)	MLR (Continuous)	WLSMV (Categorical)
<i>Corr</i> (SFT1, SFT2)	0.604*** (<i>N</i> = 1,637)	0.634*** (<i>N</i> = 1,637)	0.568*** (<i>N</i> = 919)	0.611*** (<i>N</i> = 919)
<i>Corr</i> (IST1, IST2)	0.526*** (<i>N</i> = 1,689)	0.606*** (<i>N</i> = 1,689)	0.475*** (<i>N</i> = 940)	0.588*** (<i>N</i> = 940)
<i>Corr</i> (MST1, MST2)	0.892*** (<i>N</i> = 1,738)	Did not converge	0.891*** (<i>N</i> = 955)	0.986*** (<i>N</i> = 955)
<i>Corr</i> (GST1, GST2)	0.677*** (<i>N</i> = 1,702)	0.792*** (<i>N</i> = 1,702)	0.676*** (<i>N</i> = 942)	0.782*** (<i>N</i> = 942)

* $p < .05$, ** $p < .01$, *** $p < .001$

Note : In the case of SFT, a longitudinal higher-order CFA in which items were treated as categorical was estimated with WLSMV.

Table S9. Tests of Measurement Invariance of SFT by Gender, Race, Education, and Age-Groups in Study 2

	<i>df</i>	χ^2	SRMR	RMSEA	CFI	Δdf	$\Delta\chi^2$	Δ SRMR	Δ RMSEA	Δ CFI
<i>Gender (N = 1560)</i>										
1. Configural Invariance	18	35.218*	.009	.035	.993					
1 versus 2						5	12.721*	.013	.001	-.002
2. Metric Invariance	23	46.158*	.022	.036	.991					
2 versus 3						5	15.336*	.000	.001	-.003
3. Scalar Invariance	28	57.935*	.022	.037	.988					
<i>Race (N = 1563)</i>										
1. Configural Invariance	18	41.886*	.010	.041	.992					
1 versus 2						5	5.277	.004	-.001	-.002
2. Metric Invariance	23	51.207*	.014	.040	.990					
2 versus 3						5	4.735	.001	-.002	-.001
3. Scalar Invariance	28	60.078*	.015	.038	.989					
<i>Education (N = 1559)</i>										
1. Configural Invariance	36	59.111*	.013	.041	.991					
1 versus 2						15	26.375*	.016	.000	-.004
2. Metric Invariance	51	84.457*	.029	.041	.987					
2 versus 3						15	23.622	-.001	.000	-.004
3. Scalar Invariance	66	108.882*	.028	.041	.983					
<i>Age (N = 1563)</i>										
1. Configural Invariance	36	51.595*	.012	.033	.994					
1 versus 2						15	17.726	.013	-.001	-.002
2. Metric Invariance	51	71.336*	.025	.032	.992					
2 versus 3						15	8.828	.000	-.003	-.001
3. Scalar Invariance	66	87.775*	.025	.029	.991					

Note : SRMR = standardized root mean squared residual; RMSEA = root mean squared error of approximation; CFI = comparative fit index; $\Delta\chi^2$ = Satorra-Bentler scaled χ^2 difference test; * $p < .05$

Table S10. Tests of Measurement Invariance of IST by Gender, Race, Education, and Age-Groups in Study 2

	<i>df</i>	χ^2	SRMR	RMSEA	CFI	Δdf	$\Delta\chi^2$	Δ SRMR	Δ RMSEA	Δ CFI
<i>Gender (N = 1645)</i>										
1. Configural Invariance	4	9.846*	.012	.042	.997					
1 versus 2						3	10.308*	.013	.004	-.003
2. Metric Invariance	7	19.577*	.025	.046	.994					
2 versus 3						3	6.550	.001	-.001	-.002
3. Scalar Invariance	10	26.683*	.026	.045	.992					
<i>Race (N = 1649)</i>										
1. Configural Invariance	4	6.480	.008	.027	.999					
1 versus 2						3	2.353	.006	-.007	.000
2. Metric Invariance	7	9.355	.014	.020	.999					
2 versus 3						3	14.286*	.008	.014	-.004
3. Scalar Invariance	10	19.344	.022	.034	.995					
<i>Education (N = 1645)</i>										
1. Configural Invariance	8	7.046	.009	.000	1.000					
1 versus 2						9	7.937	.013	.000	.000
2. Metric Invariance	17	14.981	.022	.000	1.000					
2 versus 3						9	8.769	.002	.000	.000
3. Scalar Invariance	26	23.426	.024	.000	1.000					
<i>Age (N = 1649)</i>										
1. Configural Invariance	8	10.003	.009	.025	.999					
1 versus 2						9	5.657	.011	-.021	.001
2. Metric Invariance	17	17.097	.020	.004	1.000					
2 versus 3						9	20.806*	.008	.023	-.004
3. Scalar Invariance	26	33.553	.028	.027	.996					

Note : SRMR = standardized root mean squared residual; RMSEA = root mean squared error of approximation; CFI = comparative fit index; $\Delta\chi^2$ = Satorra-Bentler scaled χ^2 difference test; * $p < .05$

Table S11. Tests of Measurement Invariance of MST by Gender, Race, Education, and Age-Groups in Study 2

	<i>df</i>	χ^2	SRMR	RMSEA	CFI	Δdf	$\Delta\chi^2$	Δ SRMR	Δ RMSEA	Δ CFI
<i>Gender (N = 1726)</i>										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						2	0.561	.006	.000	.000
2. Metric Invariance	2	0.561	.006	.000	1.000					
2 versus 3						2	2.946	.006	.000	.000
3. Scalar Invariance	4	3.719	.012	.000	1.000					
<i>Race (N = 1730)</i>										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						2	0.950	.007	.000	.000
2. Metric Invariance	2	0.950	.007	.000	1.000					
2 versus 3						2	1.854	.005	.000	.000
3. Scalar Invariance	4	2.823	.012	.000	1.000					
<i>Education (N = 1727)</i>										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						6	11.683	.028	.047	-.008
2. Metric Invariance	6	11.683	.028	.047	.992					
2 versus 3						6	6.764	.005	-.013	-.001
3. Scalar Invariance	12	17.993	.033	.034	.991					
<i>Age (N = 1730)</i>										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						6	2.903	.014	.000	.000
2. Metric Invariance	6	2.903	.014	.000	1.000					
2 versus 3						6	19.818*	.018	.048	-.017
3. Scalar Invariance	12	23.783*	.032	.048	.983					

Note : SRMR = standardized root mean squared residual; RMSEA = root mean squared error of approximation; CFI = comparative fit index; $\Delta\chi^2$ = Satorra-Bentler scaled χ^2 difference test; * $p < .05$

Table S12. Tests of Measurement Invariance of GST by Gender, Race, Education, and Age-Groups in Study 2

	<i>df</i>	χ^2	SRMR	RMSEA	CFI	Δdf	$\Delta\chi^2$	Δ SRMR	Δ RMSEA	Δ CFI
<i>Gender (N = 1670)</i>										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						2	2.673	.017	.020	-.001
2. Metric Invariance	2	2.673	.017	.020	.999					
2 versus 3						2	11.484*	.005	.036	-.011
3. Scalar Invariance	4	14.547*	.022	.056	.988					
<i>Race (N = 1674)</i>										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						2	1.588	.012	.000	.000
2. Metric Invariance	2	1.588	.012	.000	1.000					
2 versus 3						2	4.617	.007	.027	-.003
3. Scalar Invariance	4	6.484	.019	.027	.997					
<i>Education (N = 1671)</i>										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						6	16.027	.032	.063	-.011
2. Metric Invariance	6	16.027*	.032	.063	.989					
2 versus 3						6	5.081	-.003	-.021	.001
3. Scalar Invariance	12	20.674	.029	.042	.990					
<i>Age (N = 1674)</i>										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						6	16.176*	.036	.064	-.012
2. Metric Invariance	6	16.176	.036	.064	.988					
2 versus 3						6	4.241	.000	-.025	.003
3. Scalar Invariance	12	19.558	.036	.039	.991					

Note : SRMR = standardized root mean squared residual; RMSEA = root mean squared error of approximation; CFI = comparative fit index; $\Delta\chi^2$ = Satorra-Bentler scaled χ^2 difference test; * $p < .05$

Table S13. Tests of Longitudinal Measurement Invariance of SFT, IST, MST, and GST in Study 2

	<i>df</i>	χ^2	SRMR	RMSEA	CFI	Δdf	$\Delta\chi^2$	Δ SRMR	Δ RMSEA	Δ CFI
<i>SFT</i> (<i>N</i> = 1637)										
1. Configural Invariance	47	98.615*	.017	.026	.991					
1 versus 2						5	14.339*	.005	.000	-.001
2. Metric Invariance	52	111.154*	.022	.026	.990					
2 versus 3						5	7.768	.000	.000	-.001
3. Scalar Invariance	57	120.618*	.022	.026	.989					
<i>IST</i> (<i>N</i> = 1689)										
1. Configural Invariance	15	18.035	.015	.011	.999					
1 versus 2						3	3.719	.003	.000	.000
2. Metric Invariance	18	21.752	.018	.011	.999					
2 versus 3						3	3.450	.001	.000	.000
3. Scalar Invariance	21	25.272	.019	.011	.999					
<i>MST</i> (<i>N</i> = 1738)										
1. Configural Invariance	5	10.138	.015	.024	.997					
1 versus 2						2	3.633	.003	.000	-.002
2. Metric Invariance	7	13.833	.018	.024	.995					
2 versus 3						2	1.854	.000	-.004	.001
3. Scalar Invariance	9	15.428	.018	.020	.996					
<i>GST</i> (<i>N</i> = 1702)										
1. Configural Invariance	5	11.911*	.016	.028	.996					
1 versus 2						2	1.517	-.003	-.004	.000
2. Metric Invariance	7	13.669	.017	.024	.996					
2 versus 3						2	4.294	.002	.000	-.001
3. Scalar Invariance	9	17.933*	.019	.024	.995					

Note : SRMR = standardized root mean squared residual; RMSEA = root mean squared error of approximation; CFI = comparative fit index; $\Delta\chi^2$ = Satorra-Bentler scaled χ^2 difference test; * $p < .05$

4. Study 2 – Robustness

Table S14. Tests of Measurement Invariance of Higher-Order SFT by Gender, Race, Education, and Age-Groups in Study 2

	<i>df</i>	χ^2	SRMR	RMSEA	CFI	Δdf	$\Delta\chi^2$	Δ SRMR	Δ RMSEA	Δ CFI
<i>Gender (N = 1560)</i>										
1. Configural	498	3861.990*	.164	.093	.797					
1 versus 2						18	16.199	.001	-.001	-.004
2. First-Order Metric	516	3945.389*	.165	.092	.793					
2 versus 3						5	13.268*	.002	.000	-.001
3. First-and-Second-Order Metric	521	3968.781*	.167	.092	.792					
3 versus 4						18	29.682	.002	.000	-.005
4. First-Order Scalar	539	4065.318*	.169	.092	.787					
4 versus 5						5	12.846*	.000	-.001	-.001
5. First-and-Second-Order Scalar	544	3094.926*	.169	.091	.786					
<i>Race (N = 1563)</i>										
1. Configural	498	4186.181*	.113	.097	.797					
1 versus 2						18	18.006	.002	.000	-.004
2. First-Order Metric	516	4274.615*	.115	.097	.793					
2 versus 3						5	2.752	.000	-.001	-.001
3. First-and-Second-Order Metric	521	4294.792*	.115	.096	.792					
3 versus 4						18	49.425*	.006	.000	-.005
4. First-Order Scalar	539	4404.485*	.121	.096	.787					
4 versus 5						5	9.974	-.001	.000	-.001
5. First-and-Second-Order Scalar	544	4437.647*	.120	.096	.786					
<i>Education (N = 1559)</i>										
1. Configural	990	4710.479*	.121	.098	.792					
1 versus 2						54	67.114	.002	-.001	-.006
2. First-Order Metric	1044	4867.351*	.123	.097	.786					
2 versus 3						15	24.438	.001	.000	-.001
3. First-and-Second-Order Metric	1059	4908.468*	.124	.097	.785					
3 versus 4						54	44.824	.000	-.001	-.006
4. First-Order Scalar	1113	5072.547*	.124	.096	.779					
4 versus 5						15	29.792*	.000	-.001	-.002
5. First-and-Second-Order Scalar	1128	5127.096*	.124	.095	.777					
<i>Age (N = 1563)</i>										
1. Configural	990	5054.614*	.173	.103	.773					
1 versus 2						54	64.834	.004	-.002	-.007
2. First-Order Metric	1044	5224.774*	.177	.101	.766					
2 versus 3						15	14.857	.001	.000	-.001
3. First-and-Second-Order Metric	1059	5256.880*	.178	.101	.765					
3 versus 4						54	129.720*	.006	-.001	-.008
4. First-Order Scalar	1113	5464.454*	.184	.100	.757					
4 versus 5						15	6.819	.000	.000	-.002
5. First-and-Second-Order Scalar	1128	5512.758*	.184	.100	.755					

Note : model identification and nested model comparisons follow guidelines outlined by Rudnev et al. (2018). SRMR = standardized root mean squared residual; RMSEA = root mean squared error of approximation; CFI = comparative fit index; $\Delta\chi^2$ = Satorra-Bentler scaled χ^2 difference test; * $p < .05$

Table S15. Tests of Measurement Invariance of IST by Gender, Race, Education, and Age-Groups in Study 2

	<i>df</i>	χ^2	SRMR	RMSEA	CFI	Δdf	$\Delta\chi^2$	Δ SRMR	Δ RMSEA	Δ CFI
<i>Gender (N = 1645)</i>										
1. Configural Invariance	4	8.442	.005	.037	1.000					
1 versus 2						4	1.386	.000	-.026	.000
2. Threshold Invariance	8	8.793	.005	.011	1.000					
2 versus 3						3	3.124	.001	-.001	.000
3. Loading Invariance	11	11.904	.006	.010	1.000					
<i>Race (N = 1649)</i>										
1. Configural Invariance	4	2.070	.002	.000	1.000					
1 versus 2						4	5.875	.004	.005	.000
2. Threshold Invariance	8	8.135	.006	.005	1.000					
2 versus 3						3	1.276	-.001	-.005	.000
3. Loading Invariance	11	9.322	.005	.000	1.000					
<i>Education (N = 1645)</i>										
1. Configural Invariance	8	7.167	.004	.000	1.000					
1 versus 2						12	23.352*	.006	.037	.000
2. Threshold Invariance	20	31.539*	.010	.037	1.000					
2 versus 3						9	5.410	.000	-.012	.000
3. Loading Invariance	29	36.567	.010	.025	1.000					
<i>Age (N = 1649)</i>										
1. Configural Invariance	8	9.289	.005	.020	1.000					
1 versus 2						12	14.526	.003	.001	.000
2. Threshold Invariance	20	23.799	.008	.021	1.000					
2 versus 3						9	4.296	.001	-.021	.000
3. Loading Invariance	29	28.557	.009	.000	1.000					

Note : model identification and nested model comparisons follow guidelines outlined by Wu and Estabrook (2016) and Svetina et al. (2020). Indicators treated as categorical and models estimated with mean-and-variance adjusted weighted least squares (WLSMV) using delta parameterization. SRMR = standardized root mean squared residual; RMSEA = root mean squared error of approximation; CFI = comparative fit index; $\Delta\chi^2$ = two-step χ^2 difference test for WLSMV; * $p < .05$

Table S16. Tests of Measurement Invariance of MST by Gender, Race, Education, and Age-Groups in Study 2

	<i>df</i>	χ^2	SRMR	RMSEA	CFI	Δdf	$\Delta\chi^2$	Δ SRMR	Δ RMSEA	Δ CFI
<i>Gender (N = 1726)</i>										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						2	3.215	.016	.027	-.001
2. Threshold-Loading Invariance	2	3.215	.016	.027	.999					
<i>Race (N = 1721)</i>										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						2	2.744	.017	.021	-.001
2. Threshold-Loading Invariance	2	2.745	.017	.021	.999					
<i>Education (N = 1727)</i>										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						6	9.077	.030	.034	-.002
2. Threshold-Loading Invariance	6	9.078	.030	.034	.998					
<i>Age (N = 1730)</i>										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						6	16.402*	.038	.063	-.007
2. Threshold-Loading Invariance	6	16.403*	.038	.063	.993					

Note : model identification and nested model comparisons follow guidelines outlined by Wu and Estabrook (2016). Indicators treated as categorical and models estimated with mean-and-variance adjusted weighted least squares (WLSMV) using delta parameterization.

SRMR = standardized root mean squared residual; RMSEA = root mean squared error of approximation; CFI = comparative fit index;

$\Delta\chi^2$ = two-step χ^2 difference test for WLSMV; * $p < .05$

Table S17. Tests of Measurement Invariance of GST by Gender, Race, Education, and Age-Groups in Study 2

	<i>df</i>	χ^2	SRMR	RMSEA	CFI	Δdf	$\Delta\chi^2$	Δ SRMR	Δ RMSEA	Δ CFI
<i>Gender (N = 1670)</i>										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						3	2.350	.005	.000	.000
2. Threshold Invariance	3	2.350	.005	.000	1.000					
2 versus 3						2	2.014	.001	.000	.000
3. Loading Invariance	5	4.440	.006	.000	1.000					
<i>Race (N = 1674)</i>										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						3	1.668	.004	.000	.000
2. Threshold Invariance	3	1.668	.004	.000	1.000					
2 versus 3						2	1.130	.002	.000	.000
3. Loading Invariance	5	2.855	.006	.000	1.000					
<i>Education (N = 1671)</i>										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						9	8.046	.009	.000	.000
2. Threshold Invariance	9	8.046	.009	.000	1.000					
2 versus 3						6	13.750*	.007	.036	-.001
3. Loading Invariance	15	22.906	.016	.036	.999					
<i>Age (N = 1674)</i>										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						9	19.125*	.014	.052	-.001
2. Threshold Invariance	9	19.125*	.014	.052	.999					
2 versus 3						6	11.516	.008	-.003	-.001
3. Loading Invariance	15	30.292*	.022	.049	.998					

Note : model identification and nested model comparisons follow guidelines outlined by Wu and Estabrook (2016) and Svetina et al. (2020). Indicators treated as categorical and models estimated with mean-and-variance adjusted weighted least squares (WLSMV) using delta parameterization. SRMR = standardized root mean squared residual; RMSEA = root mean squared error of approximation; CFI = comparative fit index; $\Delta\chi^2$ = two-step χ^2 difference test for WLSMV; * $p < .05$

Table S18. Tests of Longitudinal Measurement Invariance of Higher-Order SFT in Study 2

	<i>df</i>	χ^2	SRMR	RMSEA	CFI	Δdf	$\Delta\chi^2$	Δ SRMR	Δ RMSEA	Δ CFI
<i>SFT</i> (<i>N</i> = 1637)										
1. Configural	1037	7384.492*	.044	.061	.805					
1 versus 2						18	9.184	.001	.000	-.001
2. First-Order Metric	1055	7441.548*	.045	.061	.804					
2 versus 3						5	18.739*	.000	.000	-.001
3. First-and-Second-Order Metric	1060	7466.357*	.045	.061	.803					
3 versus 4						18	43.295*	.001	.000	-.002
4. First-Order Scalar	1078	7551.846*	.046	.061	.801					
4 versus 5						5	7.414	.000	.000	.000
5. First-and-Second-Order Scalar	1083	7573.518*	.046	.061	.801					

Note : model identification and nested model comparisons follow guidelines outlined by Rudnev et al. (2018) but with first-order residual covariances, first-order factor variances, and second-order factors freely covarying across time. SRMR = standardized root mean squared residual; RMSEA = root mean squared error of approximation; CFI = comparative fit index; $\Delta\chi^2$ = Satorra-Bentler scaled χ^2 difference test; * $p < .05$

Table S19. Tests of Longitudinal Measurement Invariance of IST, MST, and GST in Study 2

	<i>df</i>	χ^2	SRMR	RMSEA	CFI	Δdf	$\Delta\chi^2$	Δ SRMR	Δ RMSEA	Δ CFI
<i>IST</i> (<i>N</i> = 1689)										
1. Configural Invariance	19	58.921*	.024	.035	.999					
1 versus 2						3	7.193	.000	-.002	.000
2. Loading Invariance	22	63.210*	.024	.033	.999					
2 versus 3						12	18.887	.001	-.009	.000
3. Threshold Invariance	34	65.936*	.025	.024	.999					
<i>MST</i> (<i>N</i> = 1738)										
1. Configural Invariance	9	72.652*	.049	.064	.987					
1 versus 2										
2. Loading-Threshold Invariance					Did not converge					
<i>GST</i> (<i>N</i> = 1702)										
1. Configural Invariance	8	236.452*	.054	.130	.985					
1 versus 2						2	3.244	.000	-.020	.002
2. Loading Invariance	10	216.277*	.054	.110	.987					
2 versus 3						9	47.783*	.001	.027	-.002
3. Threshold Invariance	19	243.591*	.055	.083	.985					

Note : model identification and nested model comparisons follow guidelines outlined by Liu et al. (2017). Indicators treated as categorical and models estimated with mean-and-variance adjusted weighted least squares (WLSMV) using delta parameterization. SRMR = standardized root mean squared residual; RMSEA = root mean squared error of approximation; CFI = comparative fit index; $\Delta\chi^2$ = two-step χ^2 difference test for WLSMV; * $p < .05$

Table S20. Tests of Longitudinal Measurement Invariance of SFT, IST, MST, and GST in Study 2

	<i>df</i>	χ^2	SRMR	RMSEA	CFI	Δdf	$\Delta\chi^2$	Δ SRMR	Δ RMSEA	Δ CFI
<i>SFT</i> (<i>N</i> = 919)										
1. Configural Invariance	47	76.666*	.017	.026	.992					
1 versus 2						5	5.185	.002	.000	-.001
2. Metric Invariance	52	83.198*	.019	.026	.991					
2 versus 3						5	9.758	.000	.000	-.001
3. Scalar Invariance	57	91.931*	.019	.026	.990					
<i>IST</i> (<i>N</i> = 940)										
1. Configural Invariance	15	11.437	.014	.000	1.000					
1 versus 2						3	3.925	.004	.000	.000
2. Metric Invariance	18	15.378	.018	.000	1.000					
2 versus 3						3	2.497	.000	.000	.000
3. Scalar Invariance	21	17.930	.018	.000	1.000					
<i>MST</i> (<i>N</i> = 955)										
1. Configural Invariance	5	6.311	.010	.017	.999					
1 versus 2						2	1.754	.003	-.004	.000
2. Metric Invariance	7	8.200	.013	.013	.999					
2 versus 3						2	1.675	.002	-.003	.000
3. Scalar Invariance	9	9.792	.015	.010	.999					
<i>GST</i> (<i>N</i> = 942)										
1. Configural Invariance	5	12.145*	.016	.039	.995					
1 versus 2						2	0.966	.002	-.008	.000
2. Metric Invariance	7	13.333	.018	.031	.995					
2 versus 3						2	3.724	.001	.000	-.001
3. Scalar Invariance	9	17.063*	.019	.031	.994					

Note : Listwise deletion used for panel attrition. SRMR = standardized root mean squared residual; RMSEA = root mean squared error of approximation; CFI = comparative fit index; $\Delta\chi^2$ = Satorra-Bentler scaled χ^2 difference test; * $p < .05$

5. Study 3 – MI/E

Table S21. Tests of Measurement Invariance of SFT, IST, MST, and GST across Samples (Studies 1, 2, and 3)

	<i>df</i>	χ^2	SRMR	RMSEA	CFI	Δdf	$\Delta\chi^2$	Δ SRMR	Δ RMSEA	Δ CFI
<i>SFT</i> (<i>N</i> = 3579)										
1. Configural Invariance	27	104.336*	.012	.049	.989					
1 versus 2						10	17.678	.005	-.002	-.003
2. Metric Invariance	37	134.691*	.017	.047	.986					
2 versus 3						10	30.965*	.000	.000	-.003
3. Scalar Invariance	47	168.425*	.017	.047	.983					
<i>IST</i> (<i>N</i> = 3732)										
1. Configural Invariance	6	14.479*	.007	.034	.998					
1 versus 2						6	8.767	.007	-.005	-.001
2. Metric Invariance	12	24.859*	.014	.029	.997					
2 versus 3						6	42.377*	.022	.013	-.005
3. Scalar Invariance	18	57.988*	.036	.042	.992					
<i>MST</i> (<i>N</i> = 3886)										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						4	5.306	.012	.016	-.001
2. Metric Invariance	4	5.306	.012	.016	.999					
2 versus 3						4	61.995*	.027	.060	-.041
3. Scalar Invariance	8	67.107*	.039	.076	.958					
<i>GST</i> (<i>N</i> = 3773)										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						4	4.133	.014	.005	.000
2. Metric Invariance	4	4.133	.014	.005	1.000					
2 versus 3						4	14.255*	.004	.026	-.006
3. Scalar Invariance	4	17.709*	.018	.031	.994					

Note : SRMR = standardized root mean squared residual; RMSEA = root mean squared error of approximation; CFI = comparative fit index; $\Delta\chi^2$ = Satorra-Bentler scaled χ^2 difference test; * $p < .05$

6. Study 3 – Robustness

Table S22. Tests of Measurement Invariance of IST, MST, and GST across Samples (Studies 1, 2, and 3)

	<i>df</i>	χ^2	SRMR	RMSEA	CFI	Δdf	$\Delta\chi^2$	Δ SRMR	Δ RMSEA	Δ CFI
<i>IST</i> (<i>N</i> = 3732)										
1. Configural Invariance	6	8.460	.003	.018	1.000					
1 versus 2						8	25.093*	.005	.017	.000
2. Threshold Invariance	14	34.750*	.008	.035	1.000					
2 versus 3						6	11.172	.001	-.004	.000
3. Loading Invariance	20	43.928*	.009	.031	1.000					
<i>MST</i> (<i>N</i> = 3886)										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						4	8.185	.017	.028	-.001
2. Threshold-Loading Invariance	4	8.185	.017	.028	.999					
<i>GST</i> (<i>N</i> = 3773)										
1. Configural Invariance	0	0.000	.000	.000	1.000					
1 versus 2						6	15.762*	.009	.037	-.001
2. Threshold Invariance	6	16.054*	.009	.037	.999					
2 versus 3						4	6.649	.002	-.006	.000
3. Loading Invariance	10	22.045*	.011	.031	.999					

Note : model identification and nested model comparisons follow guidelines outlined by Wu and Estabrook (2016) and Svetina et al. (2020). Indicators treated as categorical and models estimated with mean-and-variance adjusted weighted least squares (WLSMV) using delta parameterization. SRMR = standardized root mean squared residual; RMSEA = root mean squared error of approximation; CFI = comparative fit index; $\Delta\chi^2$ = two-step χ^2 difference test for WLSMV; * $p < .05$

Table S23. Predictive Validity of SFT, IST, MST, and GST in Study 3, OLS Regression

	Proportion of Tokens Sent by Player A to Player B			
	Response Propensity Weights		Unweighted	
	No Controls (<i>N</i> = 45)	Controls (<i>N</i> = 44)	No Controls (<i>N</i> = 45)	Controls (<i>N</i> = 44)
Standardized Beta (SFT)	0.390**	0.465**	0.394*	0.486**
Standardized Beta (IST)	0.135	0.045	0.129	0.095
Standardized Beta (MST)	0.262*	0.258	0.261 ^x	0.237
Standardized Beta (GST)	0.053	-0.027	.047	-0.041

	Proportion of Tokens Returned by Player B to Player A			
	Response Propensity Weights		Unweighted	
	No Controls (<i>N</i> = 35)	Controls (<i>N</i> = 34)	No Controls (<i>N</i> = 35)	Controls (<i>N</i> = 34)
Standardized Beta (SFT)	-0.034	-0.055	0.019	0.061
Standardized Beta (IST)	0.112	0.126	0.159	0.244
Standardized Beta (MST)	0.191	0.064	0.208	0.156
Standardized Beta (GST)	0.096	-0.211	0.117	-0.162

^x $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

Note : Control variables include measures of social value orientations (individualist-competitive orientation = referent), risk-seeking, betrayal aversion, gender, race-ethnicity, religious denomination, and years of education.

Table S24. Predictive Validity of SFT, IST, MST, and GST in Study 3, Fractional Regression

	Proportion of Tokens Sent by Player A to Player B			
	Response Propensity Weights		Unweighted	
	No Controls (<i>N</i> = 45)	Controls (<i>N</i> = 44)	No Controls (<i>N</i> = 45)	Controls (<i>N</i> = 44)
Log-odds ratio (SFT)	1.212**	1.661**	1.217**	1.793**
Log-odds ratio (IST)	0.463	0.377	0.413	0.509
Log-odds ratio (MST)	1.493*	1.772 ^x	1.494*	1.709 ^x
Log-odds ratio (GST)	0.268	-0.261	0.226	-0.321
	Proportion of Tokens Returned by Player B to Player A			
	Response Propensity Weights		Unweighted	
	No Controls (<i>N</i> = 35)	Controls (<i>N</i> = 34)	No Controls (<i>N</i> = 35)	Controls (<i>N</i> = 34)
Log-odds ratio (SFT)	-0.075	-0.110	0.047	0.195
Log-odds ratio (IST)	0.394	0.706	0.547	1.235 ^x
Log-odds ratio (MST)	0.891	0.416	1.035	0.916
Log-odds ratio (GST)	0.387	-0.893 ^x	0.535	-0.737

^x $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

Note : Control variables include measures of social value orientations (individualist-competitive orientation = referent), risk-seeking, betrayal aversion, gender, race-ethnicity, religious denomination, and years of education.

7. References

- Liu, Yu, Roger E. Millsap, Stephen G. West, Jenn-Yun Tein, Rika Tanaka, and Kevin J. Grimm. 2017. "Testing Measurement Invariance in Longitudinal Data with Ordered-Categorical Measures." *Psychological Methods* 22: 486-506.
- Rudnev, Maskim, Ekaterina Lytkina, Eldad Davidov, Peter Schmidt, and Andreas Zick. 2018. "Testing Measurement Invariance for a Second-Order Factor. A Cross-National Test of the Alienation Scale." *Methods, Data, Analyses* 12: 47-76.
- Svetina, Dubravka, Leslie Rutkowski, and David Rutkowski. 2020. "Multiple-Group Invariance with Categorical Outcomes using Updated Guidelines: An Illustration Using *Mplus* and the *lavaan*/*semTools* Packages." *Structural Equation Modeling* 27: 111-130.
- Wu, Hao and Ryne Estabrook. 2016. "Identification of Confirmatory Factor Analysis Models of Different Levels of Invariance for Ordered Categorical Outcomes." *Psychometrika* 81: 1014-1045.