Does Measurement Bias Explain Sex Differences in Self-Reported Empathy?

Thomas G. McCauley*

University of California, San Diego

William H.B. McAuliffe*

Rethink Priorities

Megan Mulhinch

University of California, San Diego

Eric J. Pedersen

University of Colorado, Boulder

Joanna Schug

College of William & Mary

Yohsuke Ohtsubo

University of Tokyo

Michael E. McCullough

University of California, San Diego

*T.G. McCauley and W.H.B. McAuliffe share first authorship on this paper.

This project/publication was made possible through the support of a grant from the John Templeton Foundation Grant 61539. The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of the John Templeton Foundation. This article was made possible through the support of Grant W911NF2010259 from the Army Research Institute. The views, opinions, and/or findings contained in this report (article) are those of the authors and shall not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documents.

Abstract

Self-reports of trait empathic concern suggest that females are more empathic than males, while self-reports of state empathy yield small, inconsistent sex differences. Researchers have inferred from this pattern that sex differences in empathic concern reflect measurement biases that are elicited by personality questionnaires but not reports of emotion experience in real time. However, sex differences in trait empathy need not correspond to sex differences in empathy in any one situation because trait questionnaires measure empathy at a higher level of aggregation. Moreover, the hypothesis that trait questionnaires of empathy are more biased than state measures has never been empirically tested. We conduct measurement invariance tests on eleven studies from our laboratory (*N* = 6,832). The effect of non-invariance upon observed sex differences were generally small, and we found some evidence that trait measures failed measurement invariance tests more frequently than state measures. A meta-analytic summary of the latent means revealed that females on average *do* experience more state and trait empathy than males, although there was substantial heterogeneity in effect sizes.

Introduction

Background

Researchers find that males are higher in agentic traits and females are higher in communal traits (Carlson, 1971; Feingold, 1994; Schwartz & Rubel, 2005). Because empathy motivates communal acts such as altruism (Batson, 2011), it is reasonable to predict that females exhibit more empathy than males do. Conversely, because many agentic acts are competitive and thus require disengagement from the concerns of others, one might expect males to exhibit less empathy. Sex differences may have evolutionary roots, as females express more empathy across non-human primates, infants, adolescents, and adults (Christov-Moore et al., 2014; Depow, Francis, & Inzlicht, 2021). Meta-analytic evidence regarding sex differences in empathy (Hoffman, 1977), along with systematic reviews of the literature (Rochat, 2022), point to women being more empathic than men.

However, researchers disagree about whether the reported sex differences truly exist . Some of the debate may be semantic rather than substantive: Researchers use the word "empathy" to refer to a family of related but distinct traits including mimicry, perspective taking, and vicarious distress (Cuff, Brown, Taylor, & Howat, 2016). Sex differences may differ across these constructs. Here, we focus on empathic concern, which is often defined as an affective reaction to a person in need or distress that is congruent with that person's own affective state (Batson, 2011); from this point on, the word "empathy" will refer to "empathic concern."

The empirical basis for our questioning whether sex differences in empathy exist, or if they are merely a measurement artifact instead, originated with an early meta-analysis (Eisenberg & Lennon, 1983) that reported that sex differences in empathy are method-dependent. Indirect measures of empathy such as heart rate deceleration, skin conductance, and facial expressions did not reveal sex differences, but more direct measures such as self-report empathy did. In other words, indirect methods may be more biased than direct measures. By bias, we mean that systematic differences in how empathy is measured might cause sex differences to be observed, even if those differences are caused by a feature of the measurement process. Although indirect measures may be less prone to response bias than questionnaire methods, they do not adequately differentiate between emotions of the same valence and similar levels of arousal (Mauss & Robinson, 2009). Questionnaires allow for greater specificity in the measurement of emotion, and Eisenberg and Lennon (1983) did find meta-analytic evidence that females reported greater empathy than did males on self-report questionnaires. However, they reported that the differences were much larger and consistent for trait empathy (individual differences in the degree to which one feels empathic in general) than state empathy (in-the-moment responses to specific instances of suffering).

Specifically, Eisenberg and Lennon (1983) reported that 94.1% of the trait empathy studies they analyzed found that females score significantly higher on a measure of trait empathy than did males, with females scoring higher by nearly a standard deviation (Glass' Δ = 0.99, p. 116). In contrast, for studies in which state empathy was measured, 64.3% of studies found there was no difference between males and females in state empathy (Eisenberg and Lennon, 1983, p. 110-113), although there was insufficient statistical information in the studies to report the meta-analytic difference The greater consistency of sex differences in trait empathy could be a by-product of the fact that studies that measure trait empathy tend to have larger samples, though Baez et al. (2017) found only trivial sex differences in state empathy for others' pain in a sample of more than 10,000 participants, reporting that the largest effect size was η^2 = 0.003 (Baez et al., 2017, p. 8).

Another common explanation for why trait questionnaires yield larger sex differences than more indirect measurement methods is that they tend to be transparent in what they are measuring, which could lead people to actively shape their responses in order to fulfill goals other than the goal of accurately reporting their traits. The fact that people are particularly likely to self-enhance on traits related to being a caring person (Paulhus & John, 1998) suggests that the desire to seem morally good could contaminate efforts to measure empathic traits accurately via self-report. Moreover, females are particularly motivated to appear empathic (Klein & Hodges, 2001), perhaps because they receive more status for possessing communal traits than do males (Buss et al., 2020). In contrast, state emotion questionnaires are typically administered just after ostensibly unplanned elicitors of empathy, and usually mix empathy adjectives in among a host of adjectives representing other emotions (e.g., Toi & Batson, 1982). The intended result is that participants remain unaware that the experimenter is interested in empathy in particular. In contrast, trait measures are more likely to activate impression management strategies, which could explain why trait empathy measures elicit larger sex differences. Supporting this hypothesis, Ickes, Gesn, and Graham (2000, p. 105) found meta-analytic evidence that females score higher on a self-report measure of empathic accuracy (*Cohen's d* = 0.56), but not on a coder-rated measure (*Cohen's d* = 0.04). Ickes et al. (2000) attributed differences between the self-report and objective measures to a demand effect that activated females' desire to appear more empathic upon learning that the study was about empathy. Although empathic accuracy is conceptually distinct from empathic concern, females's self-report assessments in both domains might have been distorted by a common desire to appear stereotypically empathic (see also Graham & Ickes, 1997). Conversely, males might be reluctant to self-report their empathy to appear stereotypically masculine or stoic (Prentice and Carranza, 2002).

Bias in trait measures of empathy could also be caused by a *recall bias*. Robinson and Clore (2002) argued that trait questionnaires tap semantic memory, which is influenced by stereotypes, whereas state reports rely on recent episodic memories. Barrett, Robin, Pietromonaco, and Eyssell's (1998) offered the larger role of stereotypes in semantic memory as a potential explanation for why females described themselves as more emotional than males on trait questionnaires, but did not report feeling a variety of emotions more strongly in a diary study of social interactions. Van Boven and Robinson (2012) also reported that sex differences in recalled emotional responses to saddening stimuli were larger when reliance on episodic memory was reduced using cognitive load manipulations.

In summary, there is empirical evidence that sex differences in empathy might depend on the method by which empathy is measured: Responses to trait measures (which typically show evidence of large sex differences) might inflate sex differences in empathy, compared to state measures of empathy (which are smaller in magnitude, compared to trait measures). Consequently, theories about empathy might exaggerate the degree to which males and females differ in empathy.

Should **Sex Differences in Trait and State Measures Correspond?**

To our knowledge, there have been no direct empirical demonstrations that trait but not state empathy questionnaires are sex-biased. Formal tests are necessary because state measures may evoke just as much socially desirable responding as trait measures do if people are proactive about the impression they wish to make on others. In fact, impression management concerns raise scores on both trait and state empathy measures (Sassenrath, 2020), and feelings of compassion in everyday life share a moderate correlation with IRI scores (Depow and Inzlicht, 2025)

However, two scales can be highly correlated, but this doesn't automatically mean they are measuring empathic concern in the same way for everyone, or that the items function identically across males and females. Furthermore, sex differences measured at the trait level can differ from sex differences of the same construct measured at the state level for reasons other than bias. Trait measures treat situation-specific variance as error whereas state measures treat situation-specific variance as valid (Epstein & O’Brien, 1985). This difference in measurement strategy respects the psychometric distinction between traits as stable, long-run expected values and states as transient deviations from those expectations (Steyer, Mayer, Geiser, & Cole, 2015). Put substantively, trait empathy refers to stable calibrations in the cognitive mechanisms underlying empathy that dictate how sensitive a person is to empathy-eliciting situations, all else equal. In contrast, state empathy reflects the output of computations that are tailored to the situation’s circumstances (Fleeson & Jayawickreme, 2015). It follows that so long as females are more empathic than males in the majority of situations that induce empathy, sex differences in trait empathy should arise even if certain situations nullify or even reverse the sex difference. Participants appear to keep this distinction in mind when they respond to trait questionnaires: Klein, Cosmides, Tooby, and Chance (2002) found that episodic memories of atypical behavior are activated when people make semantic generalizations, so that people view those atypical episodes as the boundary condition for the behaviors they expect. This might limit how generalizations are applied when making predictions about the future, such as predictions about how one might act in the hypothetical situations featured in trait questionnaire items. On this view, we might interpret trait differences in empathy, then, as participants’ reports of what they think they are like in general, which is not in tension with how they might react in a particular situation.

Although recall bias and impression management are separate routes by which spurious sex differences in empathy could arise, they both imply that, on average, a male and female with the same level of trait empathy might not have the same observed scores on a trait empathy measure.

Instead, the intrusion of gender stereotypes could either cause the male to lower his scores, the female to raise her score, or both. In contrast, recall bias should not affect state empathy measures because participants are relying on episodic memory to report their feelings. Impression management explanations of sex differences in empathy are also consistent with the proposal that socially desirable responding also plays a role in state empathy measures, albeit a smaller one than in trait measures. Both explanations imply that sex differences in both trait and state empathy would disappear if the measurement biases did not exist.

An alternative perspective implies that sex differences in empathy would remain even if the measurement biases did not exist. Importantly, though, genuine sex differences in empathy are compatible with the existence of biases that increase their apparent magnitude. The realist perspective is also agnostic to whether the sex differences arose through gender norm internalization, sex-specific selection pressures over evolutionary history, or both.

Here, we examine whether there are sex differences in empathy after correcting for measurement bias. We rely on all studies from our laboratory, both published and unpublished, in which we measured trait or state empathy. These studies were all conducted between 2016 and 2024. A boon of this approach is that we designed all studies with an eye towards high experimental realism. All objects of empathy were either confederates who were currently distressed, confederates who had just ostensibly been insulted, or actual victims in need depicted in charity campaign videos. Low realism may explain the trivial effect sizes reported by Baez et al. (2017), who elicited empathy by having participants watch stop-motion animated scenes in which the victim’s face was not visible, whereas people typically feel empathy for victims with identifiable features (Bloom, 2017). In contrast, the studies whose data we will consider here had a high degree of psychological realism: They are much more likely to be good models of the social-psychological conditions under which people might be most likely to experience empathy because they feature situations in which participants encountered real, or ostensibly real, victims in need of help (Batson, 2002, 2011).

To detect measurement bias, we use structural equation models, which make explicit how item responses (here, ratings on a questionnaire) respond to the underlying construct of interest (here, empathy). Item responses possess *measurement invariance* across the sexes if and only if the causal relationship between the construct and questionnaire responses do not differ on average between males and females (Meredith, 1993). Mean group differences and group-specific path coefficients produced by instruments which lack measurement invariance are indistinguishable from differences caused by other factors, such as impression management concern, recall bias, or in this case, demand-related responding (Guenole & Brown, 2014).

To our knowledge, no other research has tested sex differences in the measurement invariance of state empathy as measured by the Emotional Response Questionnaire (ERQ; Coke et al., 1978), a battery of empathy adjectives and non-empathic distractor items (e.g., angry, happy, confused) that describe empathic feelings in a single real-time episode, typically after subjects read about the plight of a person in need. Although there are a few other measures of state empathy (Heyers et al., 2025), the ERQ is among the most popular measures of state empathy: Since Coke, Batson, & McDavis (1978) introduced the ERQ, their work has been cited 1,576 times as of May 2025. However, several research teams have done so for the Interpersonal Reactivity Index (IRI; Davis, 1983 – one of the most common measures of trait empathy – mostly in the context of validating new versions of the IRI with respect to language or length. Guihard (2023) found evidence of scalar non-invariance according to the χ^2 test[1] when comparing the responses of males

[1] Although the authors reported neither the $\Delta\chi^2$ results nor their significance, we computed these values ourselves based on the data reported in Table 3 on p. 6524.

and females who completed a French version of all four subscales from the IRI, including one non-invariant intercept for an unspecified item belonging to the empathic concern subscale. Ingoglia, Lo Coco, and Albiero (2016) tested the measurement invariance of a shortened 16-item four-factor version of the IRI. They reported results for all four correlated factors, which revealed configural, metric, scalar, and strict non-invariance according to the χ^2 test, although there was no worsening of CFI ≥ .010, or or RMSEA ≥ .015. In contrast, Diotaiuti et al. (2021) also tested the short-form version of an Italian version of the IRI, and found that the four-factor model was fully invariant. Lucas-Molina et al. (2017) tested the measurement invariance of a Spanish version of the IRI, and found poor configural model fit according to χ^2, CFI, RMSEA, and SRMR indices, as well as evidence of metric and scalar non-invariance according to the χ^2 test[2]. Grevenstein (2020) tested the invariance of a German version of the IRI, and found evidence of scalar and strict non-invariance according to the χ^2 test.

Most studies, then, have at least found evidence of scalar non-invariance for the IRI. It remains unclear, then, whether measures of trait empathy are more unbiased than corresponding measures of state empathy: If one were to find evidence of more non-invariance in the trait measures, compared to the state measures, it would suggest that the trait measures are more biased than the state measures. In the present study, we compare the invariance of state and trait measures both within the same sample, and across different measures.

Finally, we note that we only included data from people who identify as male or female for three reasons. First, most studies only provided participants with the response options “male” or “female” in response to questions asking them to indicate their sex. Second, in the two studies that did include response options for non-binary (studies D and J), a total of n = 17 (0.01% of all participants included in those studies) endorsed that option, so that there wouldn’t be enough data

[2] We computed the $\Delta\chi^2$ values ourselves based on the data reported in Table 3 on p. 593.

to estimate CFA models for non-binary participants. Third, there is little theoretical basis for questions about sex differences for non-binary people vs. males or females, so it was unclear how our predictions about measurement bias might apply beyond males and females.

Method

Transparency, openness, and reproducibility

We did not pre-determine the sample sizes for the studies included here. We also did not preregister the hypotheses and analyses associated with this project, as this was a secondary data analysis effort. All data exclusions, manipulations, measures, transformation of measures relevant for each study, and information about the diversity of samples are reported in the Supplemental Materials. Readers may refer to Table 1 for a list of studies that appeared in previous publications. All data, analysis code, and research materials are available at https://osf.io/r5cxn/?view_only=61edf6d9e97f4853a6b544c275fe1625. Data were analyzed using R, version 4.0.0 (R Core Team, 2013). Packages used for analyses include: dmacs (Dueber, 2019); Lavaan, version 0.6-11 (Rosseel, 2012); semTools, version 0.5-5 (Jorgensen et al., 2018); metafor, version 4.6-0 (Viechtbauer, 2010); and psych, version 2.2.3 (Revelle & Revelle, 2015). This study's design and its analysis were not preregistered.

Measures of participant sex

For each study, participants were asked to report their sex. For all studies, we included the response options "male" and "female." Since gender and sex are dissociable concepts (Morgenroth, & Ryan, 2021), and because we did not measure the gender of the participants per se, we treat responses to "male" and "female" as participants' self-reported biological sex, and use the terms "male" and "female" throughout this paper. Please see the supplemental materials for information about subjects who were removed from analyses because they indicated they were non-binary, or chose not to report their sex.

Experiential and trait empathy measures

We evaluated the measurement invariance of items from the ERQ, which includes adjectives such as “empathic,” “compassionate,” “warm,” “tender,” and “moving.” The ERQ is typically administered following empathy-inducing stimuli, such as notes from laboratory confederates who express a state of need. To evaluate trait empathy, we tested the measurement invariance of the empathic concern subscale of the IRI. The IRI includes items such as “I am often quite touched by things that I see happen” and “Other people's misfortunes do not usually disturb me a great deal” (negatively worded).

We focused on the ERQ and IRI as validational targets because they are perhaps the most highly cited measures of trait and state empathy: As of July 2024, the 1980 paper in which Davis introduced the IRI has been cited 14,778 times, and the 1978 paper in which Coke et al. (1978) introduced their adjective-based approach to measuring empathy as an episodic state has been cited 1,521 times. Both measures are also highly correlated with other common measures of empathic concern, as Baldner and McGinley (2014) found that both the ERQ and the IRI shared moderate to large correlations with other measures thought to reflect empathic concern, including the Toronto Empathy Questionnaire (Spreng et al., 2009), Basic Empathy Scale - Affective subscale (Jolliffe & Farrington, 2006), and General Trait Sympathy (Lee, 2009).

Measurement invariance testing

Measurement invariance is based on the premise that observed scores on an instrument reflect underlying differences on a latent construct. Scores on the instrument are hypothesized to conform to a confirmatory factor analysis (CFA) model of the following form:

$$y_{ij} = \tau_i + \lambda_i \eta_{ij} + \varepsilon_{ij}$$

That is, y observed scores are a function of intercept τ, a regression coefficient λ, and a residual error term ε for the ith participant in group j. A measure is invariant between groups if and only if the relationship between respondents' observed scores on the instrument and their underlying construct is independent of their group membership (Wicherts & Dolan, 2010). Although measurement invariance is typically framed as a group discrepancy caused by differences in how different groups of people interpret the items, measurement invariance can also be caused by differences in different groups' motivations for responding. More generally, measurement invariance is not defined by differences in item interpretation, but instead merely by the statistical probability of obtaining a test score, conditional on some kind of group membership (Millsap, 2007). Indeed, past research on the influence of gender stereotypes on high-stakes testing has found that differences in motivations stemming from gender stereotypes does affect measurement invariance (Wicherts et al., 2005).

Measurement invariance testing typically begins with a "configural model," which assumes only that the factor structure underlying observed scores is equivalent between groups; this approach is referred to as the bottom-up approach (van der Veld & Saris, 2018)[3]. Zero-constraints (such as the absence of certain cross-loadings and residual covariances) must hold across groups, but parameter estimates for loadings, intercepts, and error terms are free to vary between groups. The testing procedure proceeds by iteratively constraining parameters equal across groups, typically starting with factor loadings. A likelihood ratio test reveals that this constraint significantly reduces model fit (i.e., the model fit gets "worse"), signaling that the model is not

[3] Another approach to measurement invariance testing is the so-called 'top-down' approach, where a researcher begins with a fully constrained model, and sequentially frees parameters (slopes, intercepts, etc.). However, if the model is misspecified, it is difficult to improve the model using the top-down approach (van der Veld & Saris, 2018), so we adopted the bottom-up approach in our analyses.

fully “metric invariant.” For example, *compassionate* could have a larger factor loading for females than males, implying that scores on this item are more closely related to the latent construct of empathy for females than they are for males. Parameters that violate metric invariance should be freely estimated for both females and males.

Next, the metric-invariant model is further restricted by constraining the intercepts of each item to be equal between groups. If this model fits worse than the metric invariant model, it lacks “scalar invariance”: Participants in one group may have higher observed scores than participants in another group, even when their placement on the latent continuum is equal. For example, if the item intercept for *compassionate* were higher in females than males, then females are higher in *observed* scores on that item than males, but this difference could not be fully attributed to sex differences in latent empathy.

Finally, modifying the scalar-invariant model such that error terms are also constrained equal across groups yields the “strict invariance” model. DeShon (2004) demonstrated that violations of strict invariance can bias comparisons of composite scores. Failures of strict invariance would imply that researchers need to use latent variable modeling rather than composite scores to validly compare groups.

The procedure described above was designed to evaluate the invariance of measures on interval scales. We measured trait and state empathy using rating scales with five or seven response categories, respectively, which is often too few to guarantee equal distances in the latent continuum among all pairs of adjacent options (Wu & Leung, 2017). Treating ordinal-scale items as if they were interval-scale items clouds whether non-invariance is due to substantive measurement differences across groups or failed distributional assumptions (Lubke & Muthen, 2004).

Estimation Strategy

To appropriately handle ordinal data (Svetina, Rutkowski, & Rutkowski, 2020) we estimated models using the diagonally-weighted least squares (DWLS) estimator with a scaled-shifted (SS) test statistic (Katsikatsou, Moustaki, Yang-Wallentin, & Jöreskog, 2012). We constrained the mean of the latent variables to 0, and the latent variances to 1 to identify the model (Kline, 2023).

Most models in this paper were parameterized using delta parameterization, by which the total latent-item-response variance is constrained to 1, and $\boldsymbol{\nu}$ (the vector of item thresholds) is constrained to 0 (Muthén & Asparouhov, 2002). A competing model parameterization method that could have been used is theta parameterization, which identifies the scale of the latent response variables by fixing the residual variance of the indicators to one. Delta and theta parameterization produce identical model fit and parameter estimates, but delta parameterization typically produces a more stable solution (Grimm & Liu, 2016). However, strict-invariance models can be identified only with theta parameterization because delta parameterization a priori constrains all residual variances to equality. For this reason, we identified strict invariance models using theta parameterization.

For the IRI items, we correlated the residuals between the negatively worded items. In the context of scales with positively-worded items, negatively worded items function as method factors that can distort the psychometric properties of measurement instruments. Negative-wording method factors act like stable latent variables (Motl & DiStefano, 2002), and CFA models that are fit to data with negatively-worded items have poor model fit when covariances among negatively worded items are omitted (Tomas & Oliver, 1999). To control for the potential confounding effect of negatively-worded items, we included residual correlations between those items.

Finally, we drew on a measurement invariance testing procedure that accounts for ordinal items possessing “item thresholds.” Item thresholds indicate how high a person’s standing on the trait of interest must be for it to be more probable that he or she will endorse at least a given item response category (e.g., a “5” or higher on a 1-7 Likert-type scale) rather than a lower category (e.g., a “4” or lower). Originally developed for the item response theory (IRT) framework, threshold models have also been developed for estimating structural equation models (Muthén, 1984; Reise, Widaman, & Pugh, 1993), including multi-group models (Li, 2016).

Researchers have found that constraining parameters such as factor loadings prior to constraining thresholds can bias the results of measurement invariance tests (Millsap & Yun-Tein 2004, p. 485; see also Wu & Estabrook, 2016, p. 1020). Consequently, for ordinal items, thresholds are first constrained to be equal between the two groups before testing the invariance of the factor loadings. The configural-invariant model is then compared to the nested threshold-constrained model using a likelihood ratio test. Next, loadings are constrained to be equal in addition to thresholds, and the threshold-invariant model is compared to the metric-invariant model. Then, the metric-invariant model is compared to a scalar-invariant model in which intercepts, loadings, and thresholds are constrained to be equal. Although researchers sometimes refer to intercepts and thresholds interchangeably (e.g., Sass, 2011), intercepts and thresholds are not equivalent for ordinal indicators in measurement invariance models. Instead, intercepts are specified for each indicator’s latent response, capturing the expected value of the observed variable if the mean of the factor is equal to zero, and thresholds are specified for each transition between adjacent response categories (Wu & Estabrook, 2016). Finally, the scalar-invariant model is compared to a strict invariance model in which thresholds, loadings, intercepts, and residual variances are constrained to be equal.

Where our constrained models failed, we used partial invariance testing to determine which indicators might be non-invariant across groups. Partial invariance was tested by examining the constrained model that failed with the lavTestScore() from the lavaan package (Rosseel, 2012), and partialInvariance() from the semTools package (Jorgensen et al., 2018), which provides functionality for estimating the Wald χ^2 test result for iteratively freeing parameters constrained in the failed model, and evaluating the change in model fit of the model with a freed parameter against the fit of the fully constrained failed model (Chou & Bentler, 1993).

Once we settled upon a model, we tested for group mean differences by constraining the latent intercept for males to 0 and freely estimating the latent intercept for females. Models were estimated using maximum likelihood estimation, and missing data were removed from analyses; all missing data was removed due to missing values coding for sex. We quantify how much group mean differences would have been biased by non-invariance by keeping non-invariant items constrained across the sexes (see Table 5 for state empathy, and Table 9 for trait empathy). In the Supplemental Materials, we also report non-invariance effect sizes for specific items (Nye & Drasgow, 2011).

Comparing sex differences for observed scores versus factor scores

Observed scores are the actual quantitative values that a person obtains on some measure, where the score reflects both the person's true underlying ability (i.e., their true score) and the presence of measurement error (Lord & Novick, 1968). Factor scores, on the other hand, decompose observed means into intercepts, factor loadings, and measurement error (Bollen & Lennox, 1991). Since observed scores conflate variance in the latent construct of interest with measurement error, observed means may share only a modest correlation with latent means, and poorly represent the construct of interest (McNeish & Wolf, 2020). Consequently, comparisons of means obtained via observed scores may lead to qualitatively different conclusions than comparisons of means obtained via factor scores (Steinmetz, 2013). Comparisons of observed scores have been used in virtually all studies that have yielded inferences regarding sex differences in empathy. Since factor analytic models partition true scores from measurement error, we compared the mean differences computed using observed scores against mean differences computed using factor scores to test whether measurement error drives sex differences.

Testing model assumptions

Having at least one “anchor item” (i.e., an item that passes tests of invariance from the configural model onwards) is a necessary precondition to conducting measurement invariance tests (Cheung & Rensvold, 1999; Nye & Drasgow, 2011). We selected anchors by fitting the models we wanted to test using the all-others-as-anchors (AOAA) strategy (Ankenmann, Witt, & Dunbar, 1999; Wang & Woods, 2017). The AOAA method determines which items are invariant between groups by iteratively comparing one model in which all item parameters are constrained to be equal to a second model in which the parameters for a single item are freely estimated in both groups. Items are deemed invariant if they do not worsen model fit when their parameters are constrained versus freed, with model fit assessed using a likelihood ratio test. If the AOAA procedure identified multiple invariant items, we selected the item which had the largest factor loading as the anchor.

Our AOAA procedure served the dual purpose of identifying individual items that functioned differently for males and females, serving as a de facto test of differential item functioning (DIF; Embretson & Reise, 2000). Typically, DIF analyses are conducted in the context of item response theory (IRT) models. However, we decided to implement the AOAA method in the context of factor analytic models in order to avoid introducing unnecessary nomenclature, since the parameters from factor analytic models with ordinal indicators are almost mathematically identical to the parameters from IRT models (Bolt, 2005). Models that were estimated using the AOAA method had their loadings, thresholds, and intercepts constrained equal for all items, save for the anchor item. The models were identified using the delta method, so that the mean of the latent factors was set to zero, and the variances of the latent factors set to 1.

Measurement invariance testing can yield false positives and false negatives when the configural model is misspecified for either group (French & Finch, 2011; Yuan & Bentler, 2004). We test whether loading all empathy adjectives onto a unidimensional empathy factor is valid for both sexes using the Satorra-Bentler χ^2 test statistic (Satorra and Bentler, 2001) with an alpha of .05. Testing for exact fit rather than relying on approximate fit indices (viz., RMSEA, CFI, SRMR) departs from conventional wisdom among applied researchers that, with increasing sample size, the chi-square test will reject essentially correct models for trivial reasons (Putnick & Bornstein, 2016). However, the chi-square test does not in fact become likely to reject correctly specified models with increasing sample size, and a small amount of misfit does not necessarily imply that the misspecification was also small (Hayduk, 2014). On the other hand, the validity of the likelihood ratio test is reduced when the configural model is misspecified (Yuan & Chan, 2016). Thus, in the Supplemental Materials we also provide changes in approximate fit indices. We also include a summary of divergences in the inferences that follow from the $\Delta\chi^2$ and ΔCFI approaches to measurement invariance testing.

Testing for “configural invariance” is distinct from later steps of invariance testing in that it can fail even if the true model is the same for both groups. Where the configural model demonstrated inadequate fit, we conducted a permutation test to determine whether the failure was caused by a difference in the factor structure between groups (group discrepancy) or a misspecification with the overall model (overall approximation discrepancy; Jorgensen, 2017; Jorgensen, Kite, Chen, & Short, 2018). If the model failure is due to overall approximation discrepancy, (2-*df*) Lagrange-multiplier test statistics can help determine which parameters should be simultaneously freed in both groups (Jorgensen, 2017). In the case that model failure was due to group discrepancy, one can consult (1-*df*) Lagrange-multiplier test statistics to detect which cross-loadings or residual covariances should be added to improve model fit (Saris, Satorra, & Sörbom, 1987). We inspected the Wald χ^2 test statistic associated with each parameter (Buse, 1982), and when deciding which parameters to free, we first freed parameters that had Bonferroni-corrected p-values < .05.

Modification indices are unable to estimate the improvement in fit from specifying a model with a different number of factors. We address the possibility of multidimensionality by fitting models that treat empathic concern as two emotions, sympathy and tenderness (Lishner, Batson, & Huss, 2011). Sympathy is elicited when observing a target that is currently in a state of need, while tenderness is evoked in response to a target that is vulnerable to harm. Currently needy individuals are also currently vulnerable, so acute need tends to evoke both sympathy and tenderness. In contrast, cues of chronic but not current vulnerability should only arouse tenderness (Lopez-Perez et al., 2019). Using principal component analysis on the ERQ, Niezink and colleagues (2012) found that sympathy and tenderness varied somewhat independently. However, a conceptualization in which sympathy and tenderness are part of the same emotion comes from research on an emotion that Zickfield, Schubert, Seibt, and Fiske (2017) call *kama muta*. *Kama muta* subsumes empathic concern because it is elicited not only by another's misfortune, but by any event that promotes the intensification of communal sharing relationships, such as a romantic gesture. It remains unclear whether empathic concern ought to be understood as a unitary emotion, or two distinct but related emotions.

Heterogeneity in empathy adjectives

Although each of the studies we describe here featured a model that was fit using some combination of empathy adjectives, the studies differed in (1) the empathy adjectives that participants rated, (2) the non-empathy adjectives that subjects rated, and (3) the empathy-eliciting stimuli to which subjects were exposed. For example, Study A included three empathy adjectives (*compassionate*, *empathic*, and, *sympathetic*), and Study B included all the empathy adjectives from Study A, as well as two additional empathy adjectives (*softhearted* and *tender*). Although our studies varied in the exact items used to measure state empathic concern, each of these items ought to be interchangeable with one another since latent variables are independent of their indicators used to measure them (Bollen, 1989). Nevertheless, we also ran one study expressly for this project that included all eight empathy adjectives (Study I) to ensure that our results generalize across all the items that might seek to measure state empathy. See Table 1 for information about which ERQ adjectives were included in each study, and Table S1 for a description of the empathy-eliciting stimulus used in each study.

Category collapsing procedures

We encountered some items with empty response categories for extreme responses, such that a given response category did not elicit any endorsement from members of one group. For example, none of the male subjects in Study B endorsed a "1" or a "2" for the adjective *sympathetic*, while some female subjects did. We collapsed across unendorsed response categories for both groups for items with empty response categories. Although collapsing categories can be problematic for ordinal measurement invariance testing (Rutkowski, Svetina, & Liaw, 2019), the discrepancies in how frequently groups used categories were small in all of our studies: Fewer than 3% of subjects in any one group endorsed a category that had zero responses for the other group. Details about how categories were collapsed for each study can be found in the Supplemental Materials.

Summary of research questions

The present study simultaneously tests several hypotheses related to theories about sex differences in empathy and how they arise, as well as psychometric questions concerning the validity of tools used to measure empathic concern more broadly. We summarize all research questions below.

AOAA anchor selection. Which single item should researchers use as an anchor item for testing hypotheses about sex differences in empathy? If researchers are limited to administering a single item, which item should they select?

Configural model selection and evaluation. How many latent variables underlie responses to self-report measures of empathic concern? Do the number of latent variables differ for males versus females? Does the configural model that best fits the data provide a good absolute fit (i.e., a non-significant χ^2 test statistic)?

Configural, metric, scalar, and strict invariance tests. Are state and trait measures of empathy biased with respect to sex, or are sex comparisons valid?

Comparison of the observed and latent means from the strict invariant model. Controlling for measurement error, do females score higher on self-report measures of empathy than men? Are there qualitative or quantitative differences for the state and trait measures' effect sizes?

Study information

Studies A-F, and I-J, included the same basic procedure for manipulating and measuring empathy (Studies G and H did not manipulate empathy). Some studies (D, I, J, and K) included both state and trait measures of empathy. For those studies, we treated the state and trait data as separate datasets, and named them accordingly. For example, studies D-1 and D-2 refer to the same dataset, but D-1 focuses on measurement invariance analyses for the state data, and D-2 focuses on analyses for the trait data.

Participants encountered information detailing how a person (or people) were suffering, or otherwise experiencing some sort of distress. Immediately after learning about the needy target, participants reported the emotions that they were currently feeling as measured by several emotion adjectives, including adjectives thought to capture empathic concern. Finally, participants were given an opportunity to help the needy target, and reported their demographic information, including their sex (i.e., whether the participant was male or female). In some studies, participants were provided with the option to respond “prefer not to answer,” or “non-binary.” Participants who responded to options other than “male” or “female,” or who didn’t provide any answer, were removed from analyses. For the state empathy analyses, we also removed participants who expressed suspicion that their partner was a sham, or stated their awareness of the experimental ruse. Participants in Studies G and H did not encounter needy targets, but they were provided with an opportunity to behave prosocially towards another person. Table 1 includes information about each study, and detailed descriptions about each study’s protocol are provided in the Supplemental Materials.

For all analyses, males were always coded as the reference group, and females were the focal group (*male* = 0, *female* = 1). The most notable difference between the studies were the adjectives used to measure empathic concern. We were restricted to testing the multidimensional hypothesis in those studies that included a sufficient number of items to identify a multi-group CFA model. Last, results for anchor analyses, permutation tests, modification indices, and AFIs for measurement invariance testing are reported in the Supplemental Materials.

Table 1

Information About Each Study Included in Analyses.

Study	*Sample (females)*	*Plat-form*	*Mean age (SD)*	*Target of empathy*	*IRI included?*	*Published?*
A	466 (240)	Lab	19.08 (0.98)	Another (sham) participant who was treated unfairly by a third party.	No	[REDACTED FOR REVIEW]
B	159 (104)	Lab	Not collected	Another (sham) participant who was experiencing a chronic financial need.	No	[REDACTED FOR REVIEW]
C	2,207 (1,126)	MTurk	36 (11.29)	Victims of Hurricane Irma and Harvey.	No	[REDACTED FOR REVIEW]
D	1,147 (639)	Lab	N/A[4]	Another (sham) participant who was experiencing emotional distress.	Yes	[REDACTED FOR REVIEW]
E	228 (154)	Lab	Not collected	Another (sham) participant who was experiencing emotional distress.	No	[REDACTED FOR REVIEW]
F	802 (405)	MTurk	36.59 (11.15)	Statistical victims in need of aid.	No	[REDACTED FOR REVIEW]
G	168 (100)	Lab	Not collected	N/A	Yes	[REDACTED FOR REVIEW]

[4] Age data from Study D was collected by having participants indicate the age bin that reflected their age (e.g., 18 – 24, 25 – 34). Since the age data do not contain single values for age, we cannot compute means and standard deviations. See the Supplemental Materials for the age bins for Study D.

H	158 (86)	Lab	Not collected	N/A	Yes	[REDACTED FOR REVIEW]
I	812 (403)	Prolific	37.85 (13.09)	A real person selected from a charity navigator.	Yes	[REDACTED FOR REVIEW]
J	295 (221)	Lab	20.51 (2.13)	People in need of help shown in effective altruism video	Yes	[REDACTED FOR REVIEW]
K	390 (215)	Prolific	35.94 (12.46)	A real person selected from the charity navigator HandUp.org.	Yes	[REDACTED FOR REVIEW]

Note: Lab = Study was conducted in the laboratory with undergraduate participants. MTurk = Study was conducted online using the Mechanical Turk research platform. Prolific = Study was conducted online using the Prolific research platform.

Results

State Empathic Concern Results

AOAA Anchor Selection

We conducted anchor analyses for the state empathy items with respect to both the unidimensional and multidimensional model specifications. Results are shown in Tables S2, S4, S6, S9, S15, S17, S25, S31, and S36. For the unidimensional model, the item *sympathetic* was selected as the anchor item in four datasets. For the multidimensional model, the item *softhearted* was selected as the anchor item in five datasets, and *sympathetic* was selected as the anchor item in four datasets. We found three instances of item non-invariance: *Concerned* (study E), *compassionate* (study I-1), and *softhearted* (study I-1).

Configural Model Selection

Before conducting measurement invariance analyses, we tested the hypothesis that empathy is best conceptualized as a single latent variable that's responsive to perceptions of a target's neediness (i.e., a unidimensional measurement model), against the hypothesis that empathy is caused by distinguishable perceptions of the neediness of a target and the vulnerability of a target (a multidimensional measurement model). We fit competing multi-group CFA models to statistically test these hypotheses. More specifically, we compared the model fit for the unidimensional model, where all items loaded onto a single factor, against the model fit of the multidimensional model, where responses loaded onto correlated sympathy and tenderness factors.

Since a minimum of two items are necessary to identify each latent factor in a correlated-factors model, we were restricted to fitting the multidimensional models in datasets that included at least two items tapping sympathy, and two items tapping tenderness; for the remainder of this paper, we term the model with latent sympathy and tenderness factors the multidimensional model. Six of the studies (B, C, D-1, F, I-1, and J-1) included a sufficient number of sympathy and tenderness items to test the multidimensional model. Table 2 details the adjectives used in each study, and whether a given adjective corresponds to sympathy or tenderness. For each study, we adjudicated between the models by conducting likelihood ratio tests that compared the model fit for the one- and multidimensional models according to a Satorra-Bentler-corrected χ^2 test, and retained the best-fitting model as the configural model for as the configural model in measurement invariance testing.

Results for the likelihood ratio tests comparing the unidimensional and multidimensional models are shown in Table 3. The multidimensional model was a significantly better fit for the data in four out of the six studies (C, D-1, and F; *p*s < .001). Although the multidimensional model was not a better fit for the data in Study B, its model fit was marginally better (p = .053). In general, this pattern of results suggests that a model which distinguishes need and vulnerability as the latent variables underlying state empathic concern, compared to a model which assumes a single latent variable.

Table 2

State Empathic Concern Adjectives Included in Each Study.

	Sympathy				**Tenderness**			
Study	*Compassionate*	*Sympathetic*	*Empathic*	*Moved*	**Concerned**	**Softhearted**	**Tender**	**Warm**
A	✓	✓	✓					
B	✓	✓		✓		✓	✓	✓
C	✓	✓	✓			✓	✓	
D-1	✓	✓	✓		✓		✓	
E	✓	✓	✓		✓			
F	✓	✓	✓			✓	✓	
I-1	✓	✓		✓		✓	✓	
J-1	✓	✓		✓		✓	✓	
K-1	✓	✓	✓	✓	✓	✓	✓	✓

Configural Model Evaluation

We retained the best-fitting model from each study as the configural model. Results are reported in Table 3. We found that the configural model failed for five out of the seven studies (C, D-1, F, I-1, and K-1), regardless of whether the best-fitting model was unidimensional or multidimensional. We conducted a permutation test on each of the failed configural models to determine if the model failed due to a discrepancy in the baseline model, or a discrepancy in the model fit for males vs. females (see Supplemental Materials for details about permutation testing and results). For studies C, F, I-1, and K-1, configural model failures were caused by a discrepancy in the baseline model (*p*s > .096), indicating that the configural failure was caused by a global misspecification that was shared by both males and females' responses, rather than a difference in the zero constraints or number of latent variables underlying males and females' responses. Since the configural model failure in studies C, F, I-1, and K-1 was due to a global misspecification, we retained the configural model for further measurement invariance testing. For Study D-1, however, we found that the configural model failure was caused by a discrepancy in the misspecification for males versus females (*Scaled* $\chi^2(32) = 171.099$, $p < .001$), suggesting that males and females differed in the number of latent variables underlying their responses to items about empathic concern. Although we cautiously proceeded with measurement invariance results for study D-1, we note that the results for this model may be biased by the different measurement models that underlie male's versus females' responses.

The multidimensional model had the best fit in five out of the seven studies (C, D-1, F, I-1, and K-1), compared to the fit of the unidimensional model. Notably, these were the same five studies where the best-fitting model failed configural model evaluation. The correlations between the latent factors all exceeded 0.85, even for those studies in which the unidimensional model provided the best fit to the data.

Table 3

Results of Configural Model Testing for State Empathic Concern.

Study	*Unidimensional model,* $\chi^2(df), p$	*Multidimensional model,* $\chi^2(df), p$	*Correlation between sympathy and tenderness,* *b (SE), p*	*Likelihood Ratio Test,* $\chi^2(df), p$
B	$X^2 = 40.822$ (38), $p = .347$	$X^2 = 27.508$ (32), $p = .693$	$b = 0.84$ (0.091), $p < .001$***	$X^2 = 12.43$ (6), $p = .053$

C	$X^2 = 203.245$ (10), $p < .001$***	$X^2 =63.510$ (8), $p < .001$***	$b = 0.98$ (0.003), $p < .001$***	$X^2 = 121.53$ (2), $p < .001$***
D-1	$X^2 = 181.395$ (42), $p < .001$***	$X^2 =171.099$ (32), $p < .001$***	$b = 0.91$ (0.034), $p < .001$***	$X^2 = 23.33$ (10), $p = .01$**
F	$X^2 = 181.935$ (10), $p < .001$***	$X^2 =39.284$ (8), $p < .001$***	$b = 0.93$ (0.013), $p < .001$***	$X^2 = 72.755$ (2), $p < .001$***
I-1	$X^2 = 156.327$ (10), $p < .001$***	$X^2 = 103.367$ (8), $p < .001$***	$b = 0.96$ (0.007), $p < .001$***	$X^2 = 48.983$ (2), $p < .001$***
J-1	$X^2 = 36.308$ (26), $p = .086$	$X^2 = 28.212$ (20), $p = .104$	$b = 0.97$ (0.02), $p < .001$***	$X^2 = 8.049$ (6), $p = .235$
K-1	$X^2 = 410.270$ (40), $p < .001$***	$X^2 = 361.479$ (38), $p < .001$***	$b = 0.97$ (0.008), $p < .001$***	$X^2 = 40.36$ (2), $p < .001$***

Note: A significant *p*-value for the likelihood ratio test indicates that the multidimensional model provides a better fit for the data than the unidimensional model. χ^2 values are Satorra-Bentler corrected. *b*s can be interpreted as correlation coefficients.

Measurement Invariance Testing for State Empathy

After evaluating the configural model, we proceeded to conduct measurement invariance analyses for state empathy with respect to (in order) threshold, metric, scalar, and strict invariance. Results are reported in Table 4. We found that, for seven of the nine studies, there were no measurement invariance violations of any kind (*p*s > .063). Overall, these results suggest that state empathy measures are not biased with respect to sex. For Studies I-1 and K-1, however, we encountered both threshold (Study I-1) and strict non-invariance (I-1 and K-1); we were able to achieve partial threshold invariance, but not partial strict invariance. See the Supplemental Materials for information about the invariance violations, modification indices, model re-fitting analyses, and a summary of divergences between the $\Delta\chi^2$ and ΔCFI approaches to invariance testing.

Table 4

Results of Measurement Invariance Analyses for the State Empathy Items.

Study	*Step in invariance testing*	*Satorra-Bentler χ^2(df)*	*p-value*
A			
Measurement invariance	Configural	37.174 (36)	.415
	ΔThreshold	3.663 (9)	.932
	ΔMetric	0.149 (2)	.928
	ΔScalar	2.199 (3)	.532
	ΔStrict	1.142 (2)	.565
	Strict	45.326 (52)	.732
B			
Measurement invariance	Configural	40.822 (38)	.347
	ΔThreshold	22.449 (22)	.433
	ΔMetric	3.969 (5)	.554
	ΔScalar	3.775 (5)	.582
	ΔStrict	11.693 (6)	.069
	Strict	85.221 (76)	.220
C			
Measurement invariance	Configural	63.510 (8)	< .001***
	ΔThreshold	13.54 (16)	.633
	ΔMetric	11.59 (7)	.115
	ΔScalar	0.083 (3)	.994
	ΔStrict	5.372 (5)	.372
	Strict	53.604 (39)	.060
D-1			

Measurement invariance	Configural	171.099 (32)	< .001***
	ΔThreshold	19.985 (20)	.459
	ΔMetric	7.291 (3)	.063
	ΔScalar	1.949 (3)	.583
	ΔStrict	2.283 (5)	.809
	Strict	192.852 (63)	< .001***
E			
Measurement invariance	Configural	16.929 (16)	.390
	ΔThreshold	14.865 (14)	.388
	ΔMetric	1.606 (3)	.658
	ΔScalar	6.628 (3)	.085
	ΔStrict	7.414 (7)	.116
	Strict	57.270 (4)	.038
F			
Measurement invariance	Configural	39.284 (8)	< .001***
	ΔThreshold	13.9 (1)	.836
	ΔMetric	2.41 (3)	.492
	ΔScalar	1.750 (3)	.626
	ΔStrict	6.840 (5)	.283
	Strict	49.968 (39)	.112
I-1			
Measurement invariance	Configural	103.367 (8)	<.001***
	ΔThreshold	43.347 (20)	.002**
	ΔPartial threshold	26.804 (18)	.083
	ΔMetric	4.162 (3)	.245
	ΔScalar	7.739 (3)	.052

	ΔStrict	17.846 (5)	.003**
	ΔPartial strict	10.708 (3)	.013*
	Partial strict	119.938 (35)	<.001***
J-1			
Measurement invariance	Configural	36.308 (26)	.086
	ΔThreshold	16.37 (20)	.693
	ΔMetric	5.164 (4)	.271
	ΔScalar	2.149 (4)	.709
	ΔStrict	3.187 (5)	.671
	Strict	56.938 (59)	.552
K-1			
Measurement invariance	Configural	361.479 (38)	<.001***
	ΔThreshold	33.009 (32)	.418
	ΔMetric	7.783 (6)	.254
	ΔScalar	10.154 (6)	.118
	ΔStrict	18.827 (8)	.016*
	ΔPartial strict	14.12 (6)	.028*
	Partial strict	301.052 (88)	<.001***

Note: For the configural models, the Satorra-Bentler χ^2 value indicates the model fit for the multi-group CFA model. For all other steps in testing, the Satorra-Bentler χ^2 value indicates the $\Delta\chi^2$ due to constraining parameters to be equivalent.

Observed Versus Latent Mean Differences

We tested whether latent mean differences provided qualitatively different information than observed mean differences. For the latent mean difference, we extracted the standardized and unstandardized latent parameter estimates from the most restricted latent variable model for which we could achieve at least partial scalar invariance. The mean difference reflects the intercept for the latent factors; the intercept is fixed to zero for males, and freely estimated for females. Unstandardized estimates for the latent mean differences (i.e., the *b*s) were the unscaled latent intercept coefficients extracted from the model; the standardized estimates (i.e., the *β*s) were computed by rescaling the latent variables so that their variance was equal to one.

For the observed difference, we regressed the observed empathic concern scores on a dummy-coded variable for sex (*Males* = 0, *Females* = 1). Unstandardized estimates for the observed difference (i.e., the *b*s) reflect the unscaled regression coefficient from the regression model; the standardized estimates (i.e., the *β*s) were computed by standardizing the data so that both the predictors and the outcomes had a mean of 0 and standard deviation of 1, and then re-running the regression analysis. For those studies that indicated that a multidimensional model provided the best fit for the data, we ran two multiple regressions that separately regressed sympathy versus tenderness on the dummy code for sex.

Results are shown in Table 5. First, we found that, with respect to the latent mean differences, females endorsed significantly more empathic concern than males in five of the nine studies (C, D-1 sympathy factor, F, I-1, J-1, and K-1; *p*s < .017). In the other four studies, mean differences between males and females were non-significant, but the difference was still in the predicted direction. For seven of the nine studies (A, B, C, F, I-1, J-1, and K-1) we found that the latent and observed mean differences were qualitatively identical. For studies D-1 and E, however, we found that the two estimates disagreed. For study D-1, there was no difference between males and females in their scores on the latent tenderness factor (*p* = .063), whereas there was a significant difference for scores on the observed tenderness composite (*p* < .001). Similarly, for study E, there was no difference between males and females in their scores on the latent empathic concern factor (*p* = .114), but there was a significant difference for scores on the observed composite (*p* = .001). In general, while there were few qualitative differences with respect to the significance level indicated by the p-values for the latent and observed models, the absolute value of the effect sizes (as indicated by the standardized *β*s) were larger and more positive for all of the latent mean differences, compared to the observed differences.

Table 5

Latent and Observed Mean Differences in State Empathic Concern for males versus females.

Study	*Latent variable model*	*Latent parameter estimate (SE)*	*Latent p-value*	*Observed parameter estimate (SE)*	*Observed p-value*
A	Strict	*β* = 0.13, *SE* = 0.11, *95% CI* = [-0.39, 0.64]; *b* = 0.14, *SE* = 0.29, *95% CI* = [-0.43, 0.71]	.629	*β* = -0.003, *SE* = 0.05, *95% CI* = [-0.25, 0.23]; *b* = -0.01, *SE* = 0.12, *95% CI* = [-0.24, 0.24]	.946
B	Strict	*β* = 0.10, *SE* = 0.32, *95% CI* = [-0.53, 0.74]; *b* = 0.10, *SE* = 0.33, *95% CI* = [-0.54, 0.75]	.752	*β* = 0.10, *SE* = 0.08, *95% CI* = [-0.12, 0.62]; *b* = 0.25, *SE* = 0.10, *95% CI* = [-0.26, 0.47]	.180
C	Strict – Sympathy	*β* = 0.28, *SE* = 0.05, *95% CI* = [0.19, 0.38]; *b* = 0.28, *SE* = 0.05, *95% CI* = [0.19, 0.38]	< .001	*β* = 0.13, *SE* = 0.02, *95% CI* = [0.08, 0.17]; *b* = 0.37, *SE* = 0.06, *95% CI* = [0.25, 0.49]	< .001

	Strict – Tenderness	*β* = 0.27, *SE* = 0.05, *95% CI* = [0.18, 0.37]; *b* = 0.27, *SE* =0.05, *95% CI* = [0.18, 0.37]	< .001	*β* = 0.12, *SE* = 0.02, *95% CI* = [0.08, 0.16]; *b* = 0.42, *SE* = 0.07, *95% CI* = [0.28, 0.55]	< .001
D-1	Strict – Sympathy	*β* = 0.71, *SE* = 0.30, *95% CI* = [0.12, 1.30]; *b* = 0.76, *SE* =0.32, *95% CI* = [0.13, 1.38]	.017	*β* = 0.21, *SE* = 0.03, *95% CI* = [0.15, 0.28]; *b* = 0.51, *SE* = 0.08, *95% CI* = [0.35, 0.67]	< .001
	Strict – Tenderness	*β* = 0.58, *SE* = 0.32, *95% CI* = [-0.04, 1.20]; *b* = 0.66, *SE* =0.35, *95% CI* = [-0.04, 1.35]	.063	*β* = 0.55, *SE* = , *95% CI* = [,]; *b* = 0.19, *SE* = 0.10, *95% CI* = [,]	< .001
E	Strict	*β* = 0.45, *SE* = 0.28, *95% CI* = [-1.00, 0.11]; *b* = 0.49, *SE* = 0.31, *95% CI* = [-1.09, 0.12]	.114	*β* = 0.22, *SE* = 0.06, *95% CI* = [0.09, 0.34]; *b* = 0.49, *SE* = 0.15, *95% CI* = [0.20, 0.79]	.001
F	Strict – Sympathy	*β* = 0.59, *SE* = 0.09, *95% CI* = [0.42, 0.76]; *b* = 0.61, *SE* = 0.09, *95% CI* = [0.44, 0.78]	< .001	*β* = 0.25, *SE* = 0.03, *95% CI* = [0.19, 0.32]; *b* = 0.73, *SE* = 0.10, *95% CI* = [0.54, 0.92]	< .001
	Strict – Tenderness	*β* = 0.62, *SE* = 0.08, *95% CI* = [0.46, 0.78]; *b* = 0.62, *SE* = 0.08, *95% CI* = [0.46, 0.78]	< .001	*β* = 0.27, *SE* = 0.03, *95% CI* = [0.20, 0.34];	< .001

				b = 0.87, *SE* = 0.11, *95% CI* = [0.66, 1.09]	
I-1	Scalar – Sympathy	*β* = 0.32, *SE* = 0.08, *95% CI* = [0.16, 0.48]; *b* = 0.29, *SE* = 0.07, *95% CI* = [0.15, 0.43]	< .001	*β* = 0.13, *SE* = 0.03, *95% CI* = [0.06, 0.20]; *b* = 0.38, *SE* = 0.10, *95% CI* = [0.18, 0.59]	< .001
	Scalar – Tenderness	*β* = 0.30, *SE* = 0.08, *95% CI* = [0.13, 0.46]; *b* = 0.26, *SE* = 0.07, *95% CI* = [0.12, 0.40]	< .001	*β* = 0.12, *SE* = 0.03, *95% CI* = [0.05, 0.19]; *b* = 0.40, *SE* = 0.12, *95% CI* = [0.17, 0.64]	< .001
J-1	Strict	*β* = 0.14, *SE* = 0.24, *95% CI* = [-0.32, 0.61]; *b* = 0.15, *SE* = 0.25, *95% CI* = [-0.34, 0.65]	.548	*β* = 0.10, *SE* = 0.06, *95% CI* = [-0.02, 0.21]; *b* = 0.32, *SE* = 0.19, *95% CI* = [-0.05, 0.69]	.094
K-1	Scalar – Sympathy	*β* = 0.41, *SE* = , *95% CI* = [0.19, 0.63]; *b* = 0.38, *SE* = 0.11, *95% CI* = [0.17, 0.59]	< .001	*β* = 0.20, *SE* = 0.05, *95% CI* = [0.10, 0.30]; *b* = 0.59, *SE* = 0.15, *95% CI* = [0.30, 0.89]	< .001
	Scalar – Tenderness	*β* = 0.33, *SE* = , *95% CI* = [0.12, 0.55]; *b* = 0.33, *SE* = 0.11, *95% CI* = [0.12, 0.54]	.003	*β* = 0.15, *SE* = 0.05, *95% CI* = [0.05, 0.25]; *b* = 0.45, *SE* = 0.15, *95% CI* = [0.15, 0.75]	.003

Note: Latent parameters are standardized coefficients (*β*), unstandardized coefficients (*b*), standard errors (SE), and 95% confidence intervals (*95% CI*) from a latent variable model. Observed parameter estimates are standardized coefficients (*β*), unstandardized coefficients (*b*), and standard errors (SE), and 95% confidence intervals (*95% CI*) from a multiple regression model. For both the latent parameter estimates, positive values indicate higher scores for females compared to males.

Meta-analysis of Effect Sizes

We then conducted meta-analyses on the mean latent differences to determine whether females endorsed more empathic concern than males. We conducted three random-effects meta-analyses, one for each of the three different factors (i.e., the unidimensional, sympathy, and tenderness factors) used in the eight studies that included state empathic concern. There were four effect sizes included for each of the unidimensional, sympathy, and tenderness meta-analyses. Results are shown in Figures 1-3. There was a non-significant difference between males and females in their endorsement of unidimensional empathic concern (*Hedge's g* = 0.21, *SE* = 0.15, *Z* = 1.46, *p* = .143; *95% CI* = [-0.07, 0.50]). In contrast, the meta-analyses for the sympathy (*Hedge's g* = 0.40, *SE* = 0.08, *Z* = 5.25, *p* < .001; *95% CI* = [0.25, 0.55]) and tenderness (*Hedge's g* = 0.38, *SE* = 0.08, *Z* = 4.60, *p* < .001; *95% CI* = [0.22, 0.54]) factors both indicated that females endorsed significantly more sympathy and tenderness than males, respectively.

Figure 1

Random-Effects Meta-analysis for the State Unidimensional Factor.

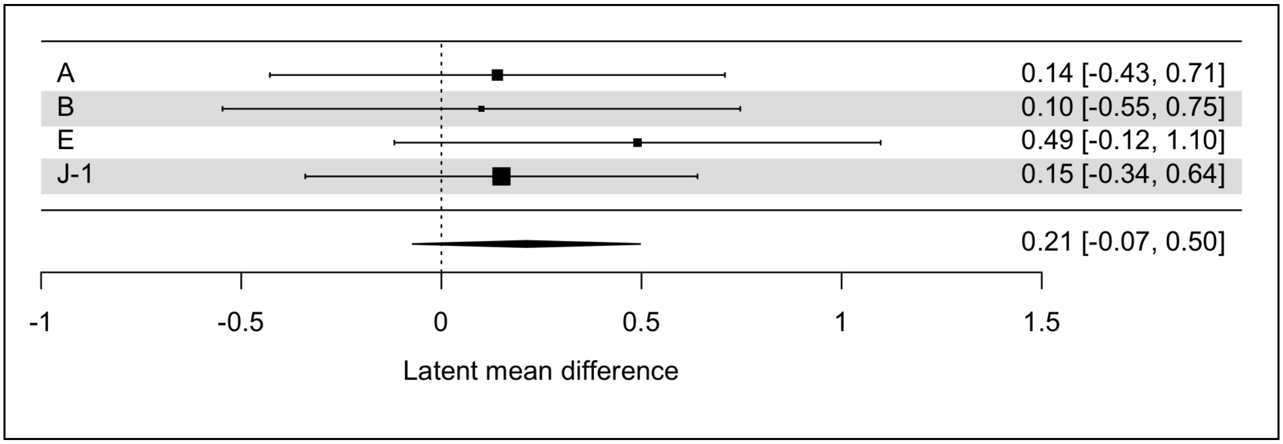


Figure 2

Random-Effects Meta-analysis for the State Sympathy Factor.

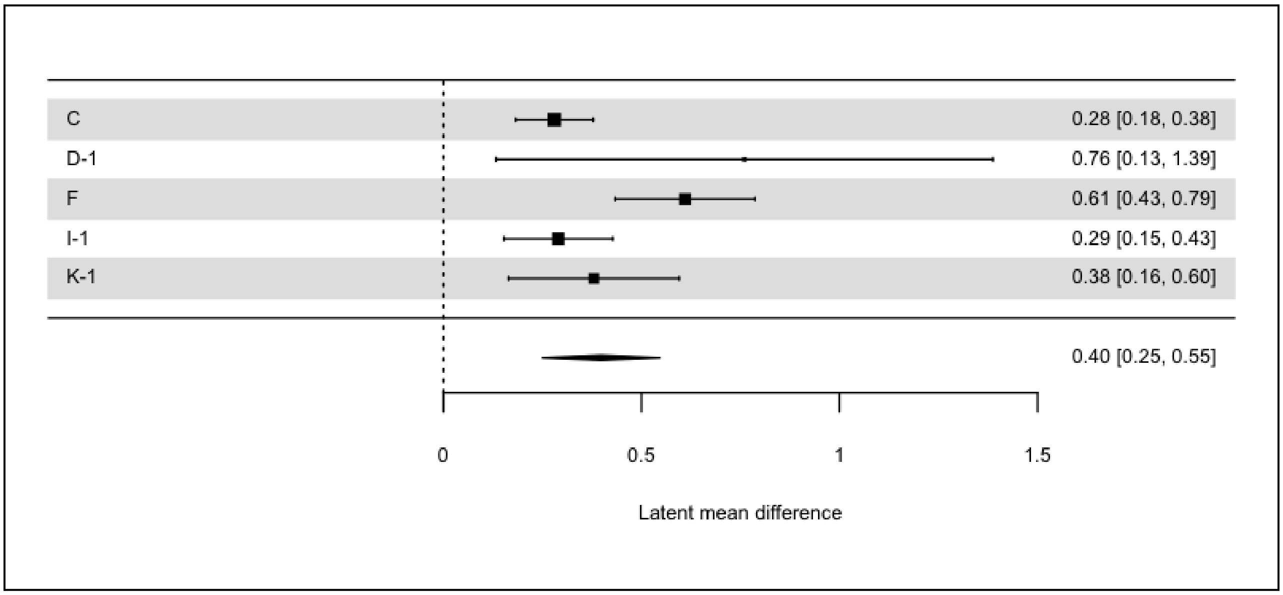
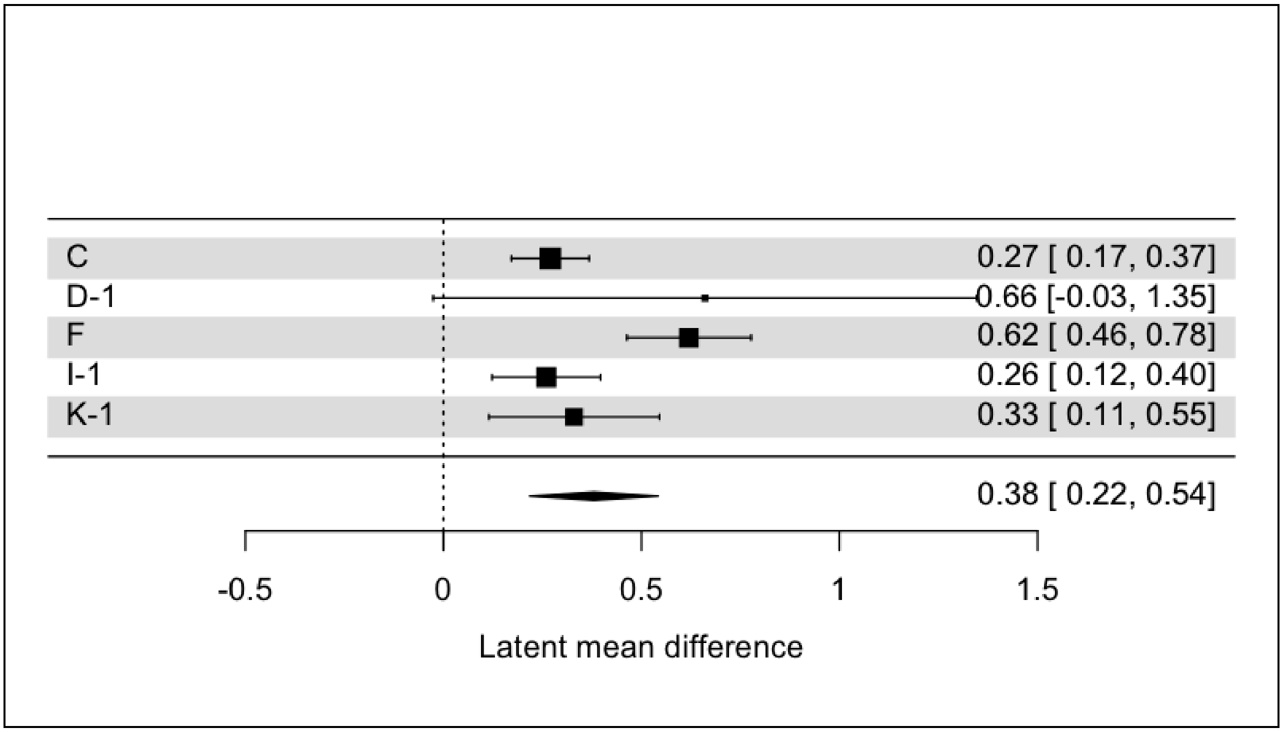


Figure 3

Random-Effects Meta-analysis for the State Tenderness Factor.



Trait Empathic Concern Results

Next, we conducted measurement invariance analyses for trait empathic concern in five studies that included the empathic concern subscale of the IRI. We followed the same procedure as we did for the state invariance testing by selecting the best-fitting configural model; evaluating the absolute fit of the configural model; conducting measurement invariance testing; and examining the latent and observed mean differences in trait empathy.

AOAA Anchor Selection

We conducted anchor analyses for the trait empathy items. Results are shown in Tables S12, S20, S23, S28, S33, and S39. The item “I often have tender, concerned feelings for people less fortunate than me” was each selected as the anchor item for five of the unidimensional models, and, for the sympathy factor, five of the multidimensional models. The item “When I see someone being taken advantage of, I feel kind of protective towards them” was also selected as an anchor item for the tenderness factor in four of the multidimensional models. We found a total of seven instances of non-invariance, with the most frequent non-invariant item being “When I see someone being treated unfairly, I sometimes don't feel very much pity for them,” which was non-invariant in two of the unidimensional models, and two of the multidimensional models.

Configural Model Selection

All six trait empathy studies included enough items to test the multidimensional model. Table 6 lists the items included in the empathic concern subscale, and the factor to which each item belongs. Results for model comparison analyses are shown in Table 7. The multidimensional model provided the best fit in four of the six studies (*p*s < .022), suggesting that, at least in some datasets, a multidimensional model that distinguishes sympathy from tenderness may provide a more accurate measurement model of empathy.

Table 6

Items From the Empathic Concern Subscale of the Interpersonal Reactivity Index, and the Factor to Which Each Item Belongs.

Sympathy	Tenderness
Sometimes I don't feel very sorry for other people when they are having problems. (R)	I often have tender, concerned feelings for people less fortunate than me.
When I see someone being taken advantage of, I feel kind of protective towards them.	I am often quite touched by things that I see happen.
Other people's misfortunes do not usually disturb me a great deal. (R)	I would describe myself as a pretty soft-hearted person.

When I see someone being treated unfairly, I sometimes don't feel very much pity for them. (R)

Note: (R) indicates an item that is reverse coded.

Configural Model Evaluation

Results are reported in Table 7. We found that the configural model was a poor fit for the data in five out of six studies (D-2, G, I-2, J-2, and K-2; *p*s < .003). Permutation tests indicated that misspecifications for all studies were caused by global discrepancies in the latent structure of trait empathic concern, rather than a discrepancy caused by different numbers of latent variables for males versus females (*p*s > .208). We retained the configural model for further measurement invariance testing for all trait empathy studies. See Supplemental Materials for details about the modifications modification indices and residual covariances.

The multidimensional model had the best fit in four out of the six studies (G, H, I-2, and K-2), compared to the fit of the unidimensional model. For the multidimensional models. the correlations between the latent factors all exceeded 0.65, even for those studies in which the unidimensional model provided the best fit to the data.

Table 7

Results of Configural Model Testing for Trait Empathic Concern.

Stu dy	*Unidimensional model, $\chi^2(df)$, p*	*Multidimensional model, $\chi^2(df)$, p*	*Correlation between sympathy and tenderness, b (SE), p*	*Likelihood Ratio Test, $\chi^2(df)$, p*
D-2	151.932 (22), *p* < .001***	152.695 (20), *p* < .001***	*b* = 0.97 (0.02), *p* < .001***	4.59 (2), *p* = .101

G	57.192 (22), $p < .001$***	41.896 (20), $p = .003$**	$b = 0.96$ (0.07), $p < .001$***	13.17 (2), $p = .001$**
H	25.132 (22), $p = .291$	17.290 (20), $p = .634$	$b = 0.65$ (0.14), $p < .001$***	7.6776 (2), $p = .022$*
I-2	65.136 (22), $p < .001$***	56.742 (20), $p < .001$***	$b = 1.02$ (0.02), $p < .001$***	7.1236 (2), p = .028*
J-2	62.849 (22), $p < .001$***	60.981 (20), $p < .001$***	$b = 0.91$ (0.06), $p < .001$***	3.347 (2), $p = .188$
K-2	84.034 (22), $p < .001$***	63.979 (20), $p < .001$***	$b = 0.88$ (0.03), $p < .001$***	14.923 (2), $p < .001$***

Note: A significant p-value for the likelihood ratio test indicates that the multidimensional model provides a better fit for the data than the one-factor model. χ^2 values are Satorra-Bentler corrected. bs can be interpreted as correlation coefficients.

Measurement Invariance Testing for Trait Empathy

Results of measurement invariance testing are summarized in Table 8. We encountered violations of measurement invariance in five out of six studies. We found evidence of threshold non-invariance in Study I-2, metric non-invariance in Study J-2, scalar non-invariance in Study G, and strict non-invariance in Studies D-2, I-2, and K-2. Overall, these results suggest that trait measures are biased with respect to sex. In particular, instances of scalar and metric non-invariance suggest that the trait measures showed at least some evidence of measurement bias. See the Supplemental Materials for information about the invariance violations, modification indices, model re-fitting analyses, and a summary of divergences between the $\Delta\chi^2$ and ΔCFI approaches to invariance testing.

Table 8

Results of Measurement Invariance Analyses for the Trait Empathy Items.

Study	*Step in invariance testing*	*Satorra-Bentler $\chi^2(df)$*	*p-value*
D-2			
	Configural	151.932 (22)	< .001***
	ΔThreshold	18.611 (14)	.180
	ΔMetric	10.132 (6)	.119
	ΔScalar	7.982 (6)	.239
	ΔStrict	28.408 (6)	< .001***
	ΔPartial strict	17.857 (4)	. 001**
	Scalar	143.912 (48)	< .001***
G			
	Configural	41.896 (20)	.003**
	ΔThreshold	18.97 (14)	.166
	ΔMetric	2.443 (5)	.785
	ΔScalar	11.40 (5)	.044*
	ΔPartial scalar	3.493 (4)	.479
	ΔStrict	0.0 (7)	.999
	Strict	63.003 (5)	.102
H			
	Configural	17.290 (20)	.634
	ΔThreshold	15.401 (12)	.220
	ΔMetric	0.871 (5)	.972
	ΔScalar	6.635 (5)	.249
	ΔStrict	11.16 (7)	.132
	Strict	50.567 (49)	.411
I-2			
	Configural	56.742 (20)	< .001***

	ΔThreshold	26.922 (14)	.020*
	ΔPartial threshold	19.304 (13)	.114
	ΔMetric	8.295 (5)	.141
	ΔScalar	6.339 (5)	.275
	ΔStrict	25.126 (7)	< .001***
	ΔPartial strict	7.212 (2)	.027
	Partial strict	103.735 (45)	< .001***
J-2	Configural	62.849 (22)	< .001***
	ΔThreshold	9.799 (13)	.710
	ΔMetric	19.076 (6)	.004**
	ΔPartial metric	2.249 (4)	.690
	ΔScalar	10.576 (6)	.102
	ΔStrict	2.323 (2)	.940
	Strict	71.628 (52)	.037*
K-2			
	Configural	63.979 (20)	< .001***
	ΔThreshold	14.796 (13)	.320
	ΔMetric	0.767 (5)	.979
	ΔScalar	8.297 (5)	.141
	ΔStrict	28.885 (7)	< .001***
	ΔPartial strict	16.506 (4)	.002**
	Scalar	71.829 (43)	.004**

Note: For the configural models, the Satorra-Bentler χ^2 value indicates the model fit for the multi-group CFA model. For all other steps in testing, the Satorra-Bentler χ^2 value indicates the $\Delta\chi^2$ due to constraining parameters to be equivalent. For the partial invariant models, Satorra-Bentler χ^2

values indicate the difference between the partial invariant model and the model from the previous invariance testing step (e.g., the partial metric invariant model is compared to the threshold invariant model).

Observed Versus Latent Mean Differences

Next, we examined the latent and observed mean differences in trait empathic concern for males versus females. Results are shown in Table 9. We found no evidence of sex differences in two of the six studies (G and H; *p*s > .271). In Studies D-2 (*b* = 0.76, *SE* = 0.08, *95% CI* = [0.60, 0.93]), I-2 ($b_{sympathy}$ = 0.57, *SE* = 0.09, *95% CI* = [0.39, 0.74]; $b_{tenderness}$ = 0.49, *SE* = 0.08, *95% CI* = [0.34, 0.64]), J-2 (*b* = 0. 95, *SE* = 0.20, *95% CI* = [0.55, 1.34]), and K-2 ($b_{sympathy}$ = *b* = 0.64, *SE* = 0.15, *95% CI* = [0.34, 0.94]; $b_{tenderness}$ = 0.66, *SE* = 0.13, *95% CI* = [0.40, 0.92]) females endorsed significantly more empathic concern than males. The results from the latent variable models were qualitatively identical to the results produced by multiple regression analyses.

Table 9

Latent and Observed Mean Differences in Trait Empathic Concern for Males Versus Females.

Study	*Latent variable model*	*Latent parameter estimate (SE)*	*Latent p-value*	*Observed parameter estimate (SE)*	*Observed p-value*
D-2	Scalar	*β* = 0.69, *SE* = 0.06, *95% CI* = [0.56, 0.82]; *b* = 0.76, *SE* = 0.08, *95% CI* = [0.60, 0.93]	< .001	*β* = 0.29, *SE* = 0.03, *95% CI* = [0.24, 0.35]; *b* = 0.46, *SE* = 0.04, *95% CI* = [0.37, 0.55]	< .001

G	Partial strict – Sympathy	*β* = -0.16, *SE* = 0.14, *95% CI* = [-0.44, 0.12]; *b* = -0.17, *SE* = 0.15, *95% CI* = [-0.46, 0.13]	.271	*β* = -0.02, *SE* = 0.06, *95% CI* = [-0.13, 0.09]; *b* = -0.03, *SE* = 0.08, *95% CI* = [-0.13, 0.20]	.687
	Partial strict – Tenderness	*β* = 0.10, *SE* = 0.14, *95% CI* = [-0.17, 0.37]; *b* = 0.10, *SE* = 0.14, *95% CI* = [-0.17, 0.37]	.473	*β* = 0.04, *SE* =0.06 , *95% CI* = [-0.07, 0.15]; *b* = 0.06, *SE* = 0.09, *95% CI* = [-0.24, 0.11]	.467
H	Strict – Sympathy	*β* = -0.06, *SE* = 0.18, *95% CI* = [-0.41, 0.30]; *b* = -0.07, *SE* = 0.23, *95% CI* = [-0.51, 0.37]	.757	*β* = -0.02, *SE* = 0.08, *95% CI* = [-0.18, 0.14]; *b* = -0.03, *SE* = 0.12, *95% CI* = [-0.21, 0.28]	.776
	Strict – Tenderness	*β* = 0.11, *SE* = 0.23, *95% CI* = [-0.33, 0.55]; *b* = 0.10, *SE* = 0.21, *95% CI* = [-0.31, 0.52]	.626	*β* = 0.06, *SE* = 0.08, *95% CI* = [-0.10, 0.22]; *b* = 0.09, *SE* = 0.12, *95% CI* = [-0.32, 0.14]	.442
I-2	Scalar – Sympathy	*β* = 0.65, *SE* = 0.12, *95% CI* = [0.41, 0.88]; *b* = 0.57, *SE* = 0.09, *95% CI* = [0.39, 0.74]	< .001	*β* = 0.32, *SE* = 0.03, *95% CI* = [0.14, 0.27]; *b* = 0.20, *SE* = 0.05, *95% CI* = [0.22, 0.43]	< .001
	Scalar – Tenderness	*β* = 0.52, *SE* = 0.09, *95% CI* = [0.34, 0.70];	< .001	*β* = 0.22, *SE* = 0.03, *95% CI* = [0.15, 0.28];	< .001

		b = 0.49, *SE* = 0.08, *95% CI* = [0.34, 0.64]		*b* = 0.39, *SE* = 0.06, *95% CI* = [0.27, 0.51]	
J-2	Partial strict	*β* = 0.75, *SE* = 0.16, *95% CI* = [0.45, 1.06]; *b* = 0.95, *SE* = 0.20, *95% CI* = [0.55, 1.34]	< .001	*β* = 0.30, *SE* = , *95% CI* = [0.19, 0.41]; *b* = 0.44, *SE* = 0.08, *95% CI* = [0.28, 0.61]	< .001
K-2	Scalar – Sympathy	*β* = 0.64, *SE* = 0.13, *95% CI* = [0.38, 0.89]; *b* = 0.64, *SE* = 0.15, *95% CI* = [0.34, 0.94]	< .001	*β* = 0.24, *SE* = 0.05, *95% CI* = [0.14, 0.33]; *b* = 0.37, *SE* = 0.08, *95% CI* = [0.22, 0.52]	< .001
	Scalar – Tenderness	*β* = 0.59, *SE* = 0.11, *95% CI* = [0.37, 0.80]; *b* = 0.66, *SE* = 0.13, *95% CI* = [0.40, 0.92]	< .001	*β* = 0.26, *SE* = 0.05, *95% CI* = [0.16, 0.36]; *b* = 0.45, *SE* = 0.08 , *95% CI* = [0.28, 0.61]	< .001

Note: Latent parameters are standardized coefficients (*β*), unstandardized coefficients (*b*), standard errors (SE), and 95% confidence intervals (*95% CI*) from a latent variable model. Observed parameter estimates are standardized coefficients (*β*), unstandardized coefficients (*b*), and standard errors (SE), and 95% confidence intervals (*95% CI*) from a multiple regression model. For both the latent parameter estimates, positive values indicate higher scores for females compared to males.

Meta-analysis of Effect Sizes

We again conducted three meta-analyses, one for each of the unidimensional, sympathy, and tenderness factors. Two studies were included in the unidimensional meta-analysis, and three studies were included in each of the sympathy and tenderness meta-analyses. Results are shown in Figures 4-6. Females endorsed significantly more unidimensional empathic concern than males (*Hedge's g* = 0.79, *SE* = 0.07, *Z* = 10.58, *p* < .001; *95% CI* = [0.64, 0.93]). There were no significant differences for sympathy (*Hedge's g* = 0.26, *SE* = 0.21, *Z* = 1.23, *p* = .217; *95% CI* = [-0.15, 0.68]), but females endorsed significantly more tenderness than males (*Hedge's g* = 0.36, *SE* = 0.14, *Z* = 2.63, *p* = .009; *95% CI* = [0.09, 0.63]) factors.

Figure 4

Random-Effects Meta-analysis for the Trait Unidimensional Factor.

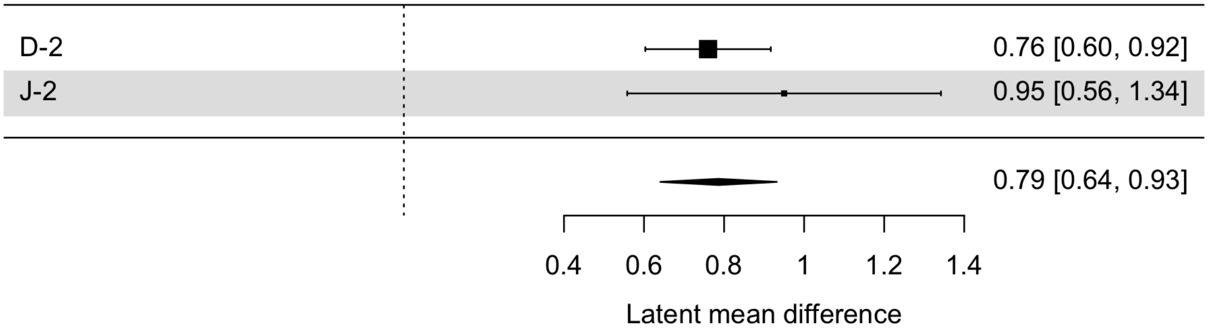


Figure 5

Random-Effects Meta-analysis for the Trait Sympathy Factor.

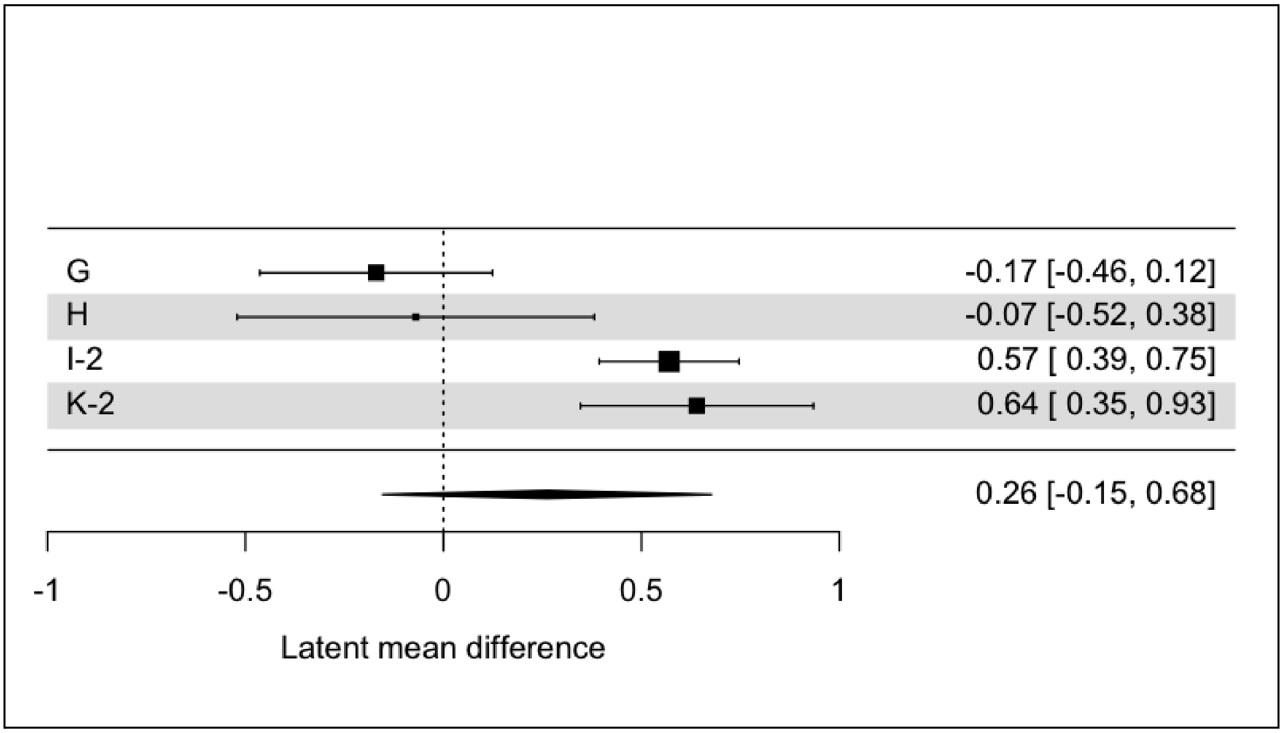
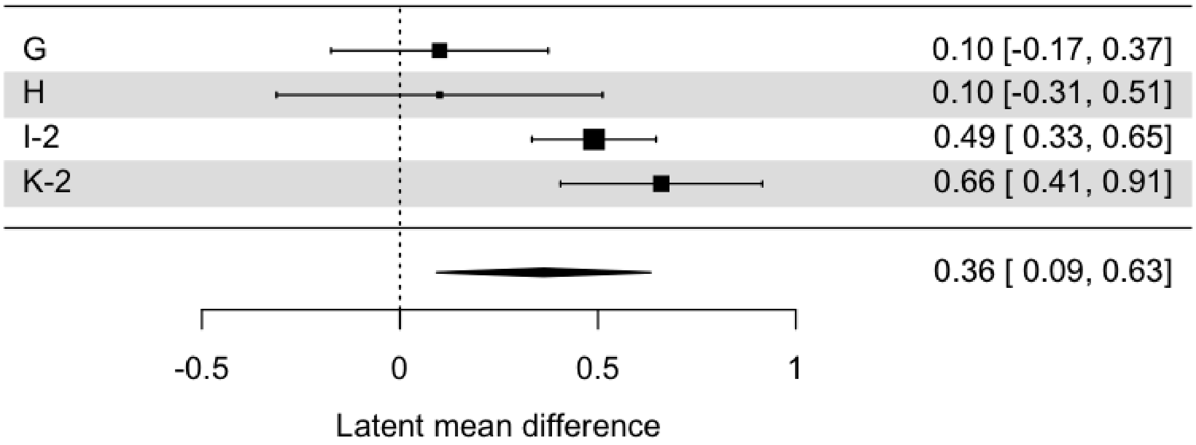


Figure 6

Random-Effects Meta-analysis for the Trait Tenderness Factor.



Discussion

Across many studies, females report higher levels of trait empathy than males. However, smaller, less consistent results for measures that assess empathy as an episodic response in real-time raises the possibility that trait-level measures exaggerate sex differences in empathy. The present paper is based on the premise that state and trait measures need not always correspond, because traits describe how people behave in general, whereas state measures reflect how people react in specific situations. Hence, the apparent lack of correspondence between the two is not evidence in itself that trait measures are biased. We investigated whether trait empathy measures are actually more biased than state empathy measures using measurement invariance testing.

We find that there is little evidence that non-invariance biases estimated sex differences in empathy at either the state or trait level. True, we did find more violations of measurement invariance for trait measures, but the latent mean differences did not materially differ from the observed mean differences. Moreover, many of the violations were either at the metric or strict level. If females were motivated to report higher trait empathy or males were motivated to report less, we would expect to see violations at the intercept or threshold level.

For both state and trait empathy measures, we found inconsistent evidence of sex differences, and the meta-analytic evidence was inconclusive with respect to whether females are more empathic than males. Still, the overall trend in meta-analytic effect sizes is consistent with the hypothesis that females experience empathy more strongly than males on average, with the smallest non-significant effect size of *Hedge's g* = 0.21. With respect to sympathy and tenderness, females endorsed significantly more sympathy and tenderness at the state level, but there was no significant difference for sympathy at the trait level (although the effect was in the predicted direction). With respect to the unidimensional models, in contrast, there was no statistically significant sex difference in empathic concern at the state level, but females endorsed significantly more empathy than males at the trait level. The finding that females are more empathic than males at the trait unidimensional level matches the findings of an experience sampling study of adults in the U.S. (Depow et al., 2021). Although the authors of this study did not conduct invariance tests, they aggregated state-level measures to test whether there are trait-level differences in empathy, so their results cannot be due to the unique biases of trait measures.

Why did we find this mixed pattern of results for the unidimensional, sympathy, and tenderness factors across the state and trait measures? We suspect the counterintuitive pattern is due to incidental differences in study characteristics between studies where a unidimensional model was preferable or the only option and studies where a two-factor model fit better. Due to the small number of studies, this spurious correlation between measurement model and study characteristics made a qualitative difference. We would predict that a larger dataset of studies would reveal more consistent results, but only future research can confirm this expectation.

Because invariance tests assume that the baseline measurement model is correct, we also had to confront ongoing debates about whether to model empathic concern as a unidimensional or multidimensional construct. Overall, we found that two-factor models fit better than one-factor models, suggesting that sympathy and tenderness are distinct traits. However, in virtually every study for which we could fit a multidimensional model to the data, the correlation between the latent sympathy and tenderness factors exceeded 0.80, even when the multidimensional model provided the better fit. On one hand, this could be interpreted as evidence that we can safely ignore the psychometric distinction between sympathy and tenderness factors, and treat responses to the ERQ and IRI as simple composite measures. On the other hand, the mega-analyses for the state and trait data indicated that sympathy and tenderness might be distinguished by their relationships with extraneous variables such as, in this case, sex. Although our initial findings here suggest that sex differences in sympathy and tenderness are similar in direction, the magnitude of their associations with sex did differ, providing at least some psychometric evidence that the inferences we draw about empathic concern and sex could be compromised by treating empathic concern as a unitary construct. In general, though, while we would still recommend that future researchers explore sex differences in sympathy and tenderness separately, it does not appear that collapsing them together has severely biased previous investigations.

A caveat is that configural model failure was a common theme even for two-factor models across studies. Hence, one might wonder if our invariance tests would have yielded different results if we had perfect knowledge about the true measurement model. Of course, one possible response here is that our use of exact fit tests is too demanding; given that the approximate fit indices suggested only minor degrees of misfit, the distorting effect on subsequent invariance tests is likely also minor. Indeed, as we report in greater detail in the supplemental materials, alternative fit indices like CFI, RMSEA, and SRMR largely indicated that the nested measurement invariance models were not significantly different from one another. Unfortunately, while minor degrees of misfit sometimes correspond to only minor misspecifications, they are also consistent with major misspecifications, so it is worth at least exploring potential misspecifications when there is a non-chance level of misfit (Ropovik, 2015). Fortunately, modification indices were not particularly consistent between studies, suggesting that the misfit may be due to sample–specific noise. Alternatively, the misfit could result from the fact that there are even more factors underlying the item covariances, but we are not aware of any theory that would posit more than two factors. One data-driven method for identifying the best-fitting model for each dataset might have been to conduct an exploratory factor analysis (EFA), or other similar methods like exploratory graph analysis (EGA; Golino & Epskamp, 2017). Ultimately, however, we decided against conducting an EFA for three reasons. First, EFA methods are insufficient for identifying the true factor structure, as EFA can identify perfect-fitting models that are still incorrectly specified (Hayduk & Glaser, 2000). Second, an EFA conducted on the same dataset as a CFA would inflate the false positive rate for the models identified via EFA. Third, the purpose of our paper was to test a theory-relevant hypothesis, not to conduct an exploratory analysis, so an EFA is beyond the scope of our efforts. However, we have made all of our data publicly available, and encourage other researchers to conduct an EFA if they so wish.

Limitations

With the exception of comparing unidimensional and multidimensional theories of empathic concern, our method for testing measurement invariance was fairly atheoretical. We did not make hypotheses about why some items would be more likely than to be biased than others with respect to testing sex differences. Instead, we used an automated tool for detecting anchor items, and a standardized statistical procedure to test for non-invariance. Although this procedure is standard in psychology, it is unclear whether its assumptions are always plausible. In particular, it presumes that at least some items are invariant, while others may be non-invariant. Many of the state empathy items under investigation here are very similar to each other, making it plausible that either all or none of them possess invariance. The methods we use here cannot distinguish between a situation where, say, all of the items violate scalar invariance from a situation where all items possess scalar invariance and there is a real group mean difference. Future researchers could use cognitive testing to try to obtain evidence that males and females are using different processes to respond to certain items (Arslan et al., 2020).

This study is also limited by the fact that none of the studies in which we measured state empathy were designed to test theories about when sex differences in empathy should be large, small, or in the opposite direction to the general trend. The situations we exposed participants to were variations on a fairly narrow theme–usually a single victim undergoing a plight for which they are blameless. Females consistently reported experiencing slightly more state empathy in these study situations. There may be a wide range of situations in which males experience more– at least no less– empathy, even if these are less common in everyday life and common experimental paradigms. For example, Bowles et al. (2024) argues that when there is little ambiguity about how one is expected to behave in a certain situation, stereotypical sex differences are smaller. Future studies could use experimental manipulations to test which situational factors affect the direction and magnitude of sex differences. If researchers could identify competing predictions based on whether sex differences exist due to norm internalization or ancestral selection pressures, then the results would help pinpoint the origin of sex differences in empathy.

Our evidence only speaks to the validity of sex differences in empathy for people who identify as male or female, and does not generalize in cases where gender doesn't match biological sex. The interactions between cultural gender prescriptions and biological sex are poorly understood, and more research is necessary to understand how gender might influence the experience and endorsement of empathy.

Our results are limited to inferences about empathic concern in particular, but many of the empathic experiences that people report in everyday life are positive (Depow et al., 2021), and other measurement instruments focusing on positive empathy may have different psychometric properties than the instruments we investigated.

Finally, our results largely reflect data collected from so-called WEIRD participants (Henrich, Heine, & Norenzayan, 2010), focusing largely on undergraduate participants from the United States. Our samples did feature a more representative sample via online platforms and community data collection, as well as participants from Japan. However, some research has investigated cross-cultural differences in empathy. While samples of participants from prototypical Western countries (e.g., United States, England) show large, significant sex differences in empathic concern, samples of participants from Asian countries (e.g., Japan, China) show small, insignificant sex differences (Kim and Lee, 2010; Zhao et al., 2018,). Zhao et al. (2022) meta-analyzed data the interaction between culture and sex in predicting empathy, and found that men from Western and Asian countries reported virtually identical trait empathy, but that women from Western countries reported significantly more empathy than women from Asian countries. In addition, Demir et al. (2022) meta-analyzed sex differences in empathy reported in dissertations submitted in Turkey found that women reported more empathy than men, providing an unbiased sample of sex differences, since none of the datasets was susceptible to publication bias. At the very least, then, there appears to be substantial cross-cultural variation in sex differences in empathy. What remains unclear, though, is whether empathy measures are measurement invariant across sexes in different cultures, and what might cause this variation. Regardless, these cross-cultural comparisons still represent just a small subset of the global population, and future research ought to extend the present research to test measurement invariance across sex participants with diverse cultural backgrounds.

Conclusion

Although our evidence does suggest that females on average probably are higher on both state and trait empathic concern than males, our present goal was less to draw conclusions about sex differences in empathy than to integrate hypotheses about whether trait emotion measures are more biased than state empathy measures into a broader psychometric framework. Because there are already numerous existing datasets with state and trait empathy measures, we encourage researchers with access to such data to re-analyze their own empathy data using the invariance procedures we used here to see whether their conclusions agree with ours.

References

Ankenmann, R. D., Witt, E. A., & Dunbar, S. B. (1999). An investigation of the power of the likelihood ratio goodness‐of‐fit statistic in detecting differential item functioning. *Journal of Educational Measurement*, *36*(4), 277-300.

Arslan, R. C., Reitz, A. K., Driebe, J. C., Gerlach, T. M., & Penke, L. (2021). Routinely randomize potential sources of measurement reactivity to estimate and adjust for biases in subjective reports. *Psychological Methods*, *26*(2), 175-185.

Baldner, C., & McGinley, J. J. (2014). Correlational and exploratory factor analyses (EFA) of commonly used empathy questionnaires: New insights. *Motivation and Emotion*, *38*, 727-744.

Baez, S., Flichtentrei, D., Prats, M., Mastandueno, R., García, A. M., Cetkovich, M., & Ibáñez, A. (2017). Men, women… who cares? A population-based study on sex differences and gender roles in empathy and moral cognition. *PloS one*, *12*(6), e0179336.

Barrett, L. F., Robin, L., Pietromonaco, P. R., & Eyssell, K. M. (1998). Are women the "more emotional" sex? Evidence from emotional experiences in social context. *Cognition & Emotion*, *12*(4), 555-578.

Batson, C. D. (2002). Addressing the altruism question experimentally. *Altruism and altruistic love: Science, philosophy, and religion in dialogue*, 89-105.

Batson, C. D. (2011). *Altruism in Humans*. Oxford University Press.

Bloom, P. (2017). Empathy and its discontents. *Trends in Cognitive Sciences, 21*(1), 24-31.

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*(2), 305-314.

Bowles, H. R., Mazei, J., & Liu, H. H. (2024). “When” Versus “Whether” Gender/Sex Differences: Insights From Psychological Research on Negotiation, Risk-Taking, and Leadership. *Perspectives on Psychological Science*, 855-873.

Buse, A. (1982). The likelihood ratio, Wald, and Lagrange multiplier tests: An expository note. *The American Statistician*, *36*(3a), 153-157.

Buss, D.M., Durkee, P.K., Shackelford, T.K., Bowdle, B.F., Schmitt, D.P., Brase, G.L., Choe, J.C., & Trofimova, I. (2020). Human status criteria: Sex differences and similarities across 14 nations. *Journal of Personality and Social Psychology*, *119*(5), 979-998.

Carlson, R. (1971). Sex differences in ego functioning: exploratory studies of agency and communion. *Journal of Consulting and Clinical Psychology*, *37*(2), 267-277.

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management*, *25*(1), 1-27.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233-255.

Chou, C. P., & Bentler, P. M. (1993). Invariant standardized estimated parameter change for model modification in covariance structure analysis. *Multivariate Behavioral Research*, *28*(1), 97-110.

Christov-Moore, L., Simpson, E. A., Coudé, G., Grigaityte, K., Iacoboni, M., & Ferrari, P. F. (2014). Empathy: Gender effects in brain and behavior. *Neuroscience & Biobehavioral Reviews*, *46*, 604-627.

Coke, J. S., Batson, C. D., & McDavis, K. (1978). Empathic mediation of helping: A two-stage model. *Journal of Personality and Social Psychology*, *36*(7), 752-766.

Cuff, B. M., Brown, S. J., Taylor, L., & Howat, D. J. (2016). Empathy: A review of the concept. *Emotion Review*, *8*(2), 144-153.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, *44*(1), 113-126.

Depow, G. J., & Inzlicht, M. (2025). How Individual Differences in Empathy Predict Moments of Empathy in Everyday Life. *Personality and Social Psychology Bulletin*, *0*(0), 1-16.

Depow, G. J., Francis, Z., & Inzlicht, M. (2021). The experience of empathy in everyday life. *Psychological Science*, *32*(8), 1198-1213.

DeShon, R. P. (2004). Measures are not invariant across groups without error variance homogeneity. *Psychology Science*, *46*, 137-149.

Diotaiuti, P., Valente, G., Mancone, S., Grambone, A., & Chirico, A. (2021). Metric goodness and measurement invariance of the italian brief version of interpersonal reactivity index: A study with young adults. *Frontiers in Psychology*, *12*, Article 773363.

Dueber, D. (2019). dmacs: Measurement nonequivalence effect size calculator. *R package version 0.1. 0 [Computer software] https://CRAN. R-project. org/package= dmacs*.

Eisenberg, N., & Lennon, R. (1983). Sex differences in empathy and related capacities. *Psychological Bulletin*, *94*(1), 100-131.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah: Lawrence Erlbaum Associates, Inc., Publishers.

Epstein, S., & O'Brien, E. J. (1985). The person–situation debate in historical and current perspective. *Psychological Bulletin*, *98*(3), 513-537.

Feingold, A. (1994). Gender differences in personality: a meta-analysis. *Psychological Bulletin*, *116*(3), 429.

Fleeson, W., & Jayawickreme, E. (2015). Whole trait theory. *Journal of Research in Personality*, *56*, 82-92.

French, B. F., & Finch, W. H. (2011). Model misspecification and invariance testing using confirmatory factor analytic procedures. *The Journal of Experimental Education*, *79*(4), 404-428.

Glass, G, V., McGaw, B,, & Smith, M. L. *Meta-analyses in social research*. Beverly Hills: Sage, 1981.

Golino, H. F., & Epskamp, S. (2017). Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PloS one*, *12*(6), Article e0174035.

Graham, T., & Ickes, W. (1997). When women's intuition isn't greater than men's. In W. Ickes (Ed.), *Empathic accuracy* (pp. 117-143). New York: Guilford Press.

Grevenstein, D. (2020). Factorial validity and measurement invariance across gender groups of the German version of the Interpersonal Reactivity Index. *Measurement Instruments for the Social Sciences, 2*(1), Article 8.

Grimm, K. J., & Liu, Y. (2016). Residual structures in growth models with ordinal outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(3), 466-475.

Guenole, N., & Brown, A. (2014). The consequences of ignoring measurement invariance for path coefficients in structural equation models. *Frontiers in Psychology*, *5*, 980.

Guihard, G. (2023). Measurement invariance analysis of two empathy scales in a sample of French first year students registered in health formation. *Current Psychology, 42*(8), 6516-6531.

Hayduk, L. (2014). Seeing perfectly fitting factor models that are causally misspecified: Understanding that close-fitting models can be worse. *Educational and Psychological Measurement*, *74*(6), 905-926.

Hayduk, L. A., & Glaser, D. N. (2000). Jiving the four-step, waltzing around factor analysis, and other serious fun. *Structural Equation Modeling, 7*(1), 1-35.

Heyers, K., Schrödter, R., Pfeifer, L. S., Ocklenburg, S., Güntürkün, O., & Stockhorst, U. (2025). (State) empathy: how context matters. *Frontiers in Psychology*, *16*, Article 1525517.

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

Hoffman, M. L. (1977). Sex differences in empathy and related behaviors. *Psychological Bulletin*, *84*(4), 712-722.

Ickes, W., Gesn, P. R., & Graham, T. (2000). Gender differences in empathic accuracy: Differential ability or differential motivation?. *Personal Relationships*, *7*(1), 95-109.

Ingoglia, S., Lo Coco, A., & Albiero, P. (2016). Development of a brief form of the Interpersonal Reactivity Index (B–IRI). *Journal of Personality Assessment, 98*(5), 461-471.

Jiang, G., Mai, Y., & Yuan, K. H. (2017). Advances in Measurement Invariance and Mean Comparison of Latent Variables: Equivalence Testing and A Projection-Based Approach. *Frontiers in Psychology*, *8*, Article 1823.

Jolliffe, D., & Farrington, D. P. (2006). Development and validation of the Basic Empathy Scale. *Journal of Adolescence*, *29*(4), 589-611.

Jorgensen, T. D. (2017). Applying permutation tests and multivariate modification indices to configurally invariant models that need respecification. *Frontiers in Psychology*, *8*, 1455.

Jorgensen, T. D., Kite, B. A., Chen, P. Y., & Short, S. D. (2018). Permutation randomization methods for testing measurement equivalence and detecting differential item functioning in multiple-group confirmatory factor analysis. *Psychological Methods*, *23*(4), 708-728.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., Rosseel, Y., Miller, P., Quick, C., & Garnier-Villarreal, M. (2018). semTools: Useful tools for structural equation modeling. *R package version 0.5*, *1*.

Katsikatsou, M., Moustaki, I., Yang-Wallentin, F., & Jöreskog, K. G. (2012). Pairwise likelihood estimation for factor analysis models with ordinal data. *Computational Statistics & Data Analysis*, *56*(12), 4243-4258.

Klein, S. B., Cosmides, L., Tooby, J., & Chance, S. (2002). Decisions and the evolution of memory: multiple systems, multiple functions. *Psychological Review*, *109*(2), 306-329.

Klein, K. J., & Hodges, S. D. (2001). Gender differences, motivation, and empathic accuracy: When it pays to understand. *Personality and Social Psychology Bulletin*, *27*(6), 720-730.

Kline, R. B. (2023). *Principles and practice of structural equation modeling*. Guilford publications.

Lee, S. A. (2009). Measuring individual differences in trait sympathy: Instrument construction and validation. *Journal of Personality Assessment*, *91*(6), 568-583.

Li, C. H. (2016). The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. *Psychological Methods*, *21*(3), 369-387.

Lishner, D. A., Batson, C. D., & Huss, E. (2011). Tenderness and sympathy: Distinct empathic emotions elicited by different forms of need. *Personality and Social Psychology Bulletin*, *37*(5), 614-625.

López-Pérez, B., Carrera, P., Oceja, L., Ambrona, T., & Stocks, E. (2019). Sympathy and tenderness as components of dispositional empathic concern: Predicting helping and caring behaviors. *Current Psychology*, *38*(2), 458-468.

Lucas-Molina, B., Pérez-Albéniz, A., Ortuño-Sierra, J., & Fonseca-Pedrero, E. (2017). Dimensional structure and measurement invariance of the Interpersonal Reactivity Index (IRI) across gender. *Psicothema, 29*(4), 590-595.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Lubke, G. H., & Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling*, *11*(4), 514-534.

Marcoulides, K. M., & Yuan, K. H. (2017). New ways to evaluate goodness of fit: A note on using equivalence testing to assess structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(1), 148-153.

Maul, A. (2017). Rethinking traditional methods of survey validation. *Measurement: Interdisciplinary Research and Perspectives*, *15*(2), 51-69.

Mauss, I. B., & Robinson, M. D. (2009). Measures of emotion: A review. *Cognition and Emotion*, *23*(2), 209-237.

McAuliffe, W. H., Forster, D. E., Philippe, J., & McCullough, M. E. (2018). Digital altruists: Resolving key questions about the empathy–altruism hypothesis in an Internet sample. *Emotion*, *18*(4), 493-506.

McCauley, T. G., & McCullough, M. E. (2022). Retrospective self-reported childhood experiences in enriched environments uniquely predict prosocial behavior and personality traits in adulthood. *Evolutionary Psychology*, *20*(3), Article 14747049221110603.

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods, 52*(6), 2287-2305.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525-543.

Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika*, *72*, 461-473.

Morgenroth, T., & Ryan, M. K. (2021). The effects of gender trouble: An integrative theoretical framework of the perpetuation and disruption of the gender/sex binary. *Perspectives on Psychological Science*, *16*(6), 1113-1142.

Motl, R. W., & DiStefano, C. (2002). Longitudinal invariance of self-esteem and method effects associated with negatively worded items. *Structural equation modeling: A Multidisciplinary Journal, 9*(4), 562-578.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*(3), 479-515.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115-132.

Muthén, B., & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. *Mplus web notes*, *4*(5), 1-22.

Niezink, L. W., Siero, F. W., Dijkstra, P., Buunk, A. P., & Barelds, D. P. (2012). Empathic concern: Distinguishing between tenderness and sympathy. *Motivation and Emotion*, *36*, 544-549.

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: understanding the practical importance of differences between groups. *Journal of Applied Psychology*, *96*(5), 966-980.

Paulhus, D. L., & John, O. P. (1998). Egoistic and moralistic biases in self‐perception: The interplay of self‐deceptive styles with basic traits and motives. *Journal of Personality*, *66*(6), 1025-1060.

Pedersen, E. J., McAuliffe, W. H., & McCullough, M. E. (2018). The unresponsive avenger: More evidence that disinterested third parties do not punish altruistically. *Journal of Experimental Psychology: General*, *147*(4), 514-544.

Prentice, D. A., & Carranza, E. (2002). What women and men should be, shouldn't be, are allowed to be, and don't have to be: The contents of prescriptive gender stereotypes. *Psychology of Women Quarterly, 26*(4), 269-281.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71-90.

R Core Team, R. (2013). R: A language and environment for statistical computing.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*(3), 552-566.

Revelle, W., & Revelle, M. W. (2015). Package ‘psych’. *The comprehensive R archive network*, *337*(338), 161-165.

Robinson, M. D., & Clore, G. L. (2002). Episodic and semantic knowledge in emotional self-report: evidence for two judgment processes. *Journal of Personality and Social Psychology*, *83*(1), 198-215.

Rochat, M. J. (2023). Sex and gender differences in the development of empathy. *Journal of Neuroscience Research*, *101*(5), 718-729.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*, 1-36.

Ropovik, I. (2015). A cautionary note on testing latent variable models. *Frontiers in Psychology*, *6*, Article 1715.

Rutkowski, L., Svetina, D., & Liaw, Y. L. (2019). Collapsing categorical variables and measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(5), 790-802.

Saris, W. E., Satorra, A., & Sörbom, D. (1987). The detection and correction of specification errors in structural equation models. *Sociological Methodology*, 105-129.

Sass, D. A. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment*, *29*(4), 347-363.

Sassenrath, C. (2020). "Let me show you how nice I am": Impression management as bias in empathic responses. *Social Psychological and Personality Science*, *11*(6), 752-760.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*(4), 507-514.

Schwartz, S. H., & Rubel, T. (2005). Sex differences in value priorities: cross-cultural and multimethod studies. *Journal of Personality and Social Psychology*, *89*(6), 1010-1028.

Spreng, R. N., McKinnon*, M. C., Mar, R. A., & Levine, B. (2009). The Toronto Empathy Questionnaire: Scale development and initial validation of a factor-analytic solution to multiple empathy measures. *Journal of Personality Assessment*, *91*(1), 62-71.

Steinmetz, H. (2013). Analyzing Observed Composite Differences Across Groups. *Methodology*, *9*(1), 1-12.

Steyer, R., Mayer, A., Geiser, C., & Cole, D. A. (2015). A theory of states and traits—Revised. *Annual Review of Clinical Psychology*, *11*(1), 71-98.

Svetina, D., Rutkowski, L., & Rutkowski, D. (2020). Multiple-group invariance with categorical outcomes using updated guidelines: an illustration using M plus and the lavaan/semtools packages. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 111-130.

Toi, M., & Batson, C. D. (1982). More evidence that empathy is a source of altruistic motivation. *Journal of Personality and Social Psychology*, *43*(2), 281.

Tomas, J. M., & Oliver, A. (1999). Rosenberg's self‐esteem scale: Two factors or method effects. *Structural Equation Modeling: A Multidisciplinary Journal,* 6(1), 84-98.

Van Boven, L., & Robinson, M. D. (2012). Boys don't cry: Cognitive load and priming increase stereotypic sex differences in emotion memory. *Journal of Experimental Social Psychology*, *48*(1), 303-309.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*, 1-67.

van der Veld, W. M., & Saris, W. E. (2018). Measurement equivalence testing 2.0. In *Cross-Cultural Analysis* (pp. 245-279). Routledge.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1-48.

Wang, M., & Woods, C. M. (2017). Anchor selection using the Wald test anchor-all-test-all procedure. *Applied Psychological Measurement*, *41*(1), 17-29.

Wicherts, J. M., & Dolan, C. V. (2010). Measurement invariance in confirmatory factor analysis: An illustration using IQ test performance of minorities. *Educational Measurement: Issues and Practice*, *29*(3), 39-47.

Wicherts, J. M., Dolan, C. V., & Hessen, D. J. (2005). Stereotype threat and group differences in test performance: a question of measurement invariance. *Journal of Personality and Social Psychology, 89*(5), 696-716.

Wu, H., & Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. *Psychometrika*, *81*(4), 1014-1045.

Wu, H., & Leung, S. O. (2017). Can Likert scales be treated as interval scales?—A simulation study. *Journal of Social Service Research*, *43*(4), 527-532.

Yuan, K. H., & Bentler, P. M. (2004). On chi-square difference and z tests in mean and covariance structure analysis when the base model is misspecified. *Educational and Psychological Measurement*, *64*(5), 737-757.

Yuan, K. H., & Chan, W. (2016). Measurement invariance via multigroup SEM: Issues and solutions with chi-square-difference tests. *Psychological Methods*, *21*(3), 405-426.

Zickfeld, J. H., Schubert, T. W., Seibt, B., & Fiske, A. P. (2017). Empathic concern is part of a more general communal emotion. *Frontiers in Psychology*, *8*, Article 723.

Supplemental Materials

Table of Contents

Study Details

Study A

Participants

We collected data from 546 participants (*females* = 279) from a mid-sized university in the southeastern United States. Participants received course credit, and could earn additional compensation based on their performance in the study. After removing 73 suspicious participants and $N = 8$ participants who had missing data, our final sample included 466 participants in data analysis (due to a computer error, descriptives for age are based on a sample of $n = 60$; $M_{age} = 19.08$, $SD_{age} = 0.98$). We did not collect additional information about diversity for this study. Data originated from Studies 2, 3, 4, and 5 reported in [REDACTED FOR REVIEW].

Procedure

The general procedure for all four experiments was as follows. Participants were informed that they would be interacting with 2-3 other participants in an essay writing and reviewing task. After writing their essays, participants learned that one of the other participants (i.e., the transgressor) insulted an essay written by another subject (the victim). Participants then self-reported their emotional reaction towards the victim. Afterwards, participants were given the opportunity to punish the transgressor.

Although the four experiments followed a similar procedure, the manipulations used in each experiment were slightly different. We controlled for these differences by regressing the latent factors on condition dummy variables, and excluding participants from conditions in which they themselves were the victim. See [REDACTED FOR REVIEW] for more details about the procedure of each experiment, and Supplementary Materials for details about the condition dummy variables.

Measures

ERQ adjectives. Empathy was measured with three empathy adjectives: “compassionate”, “empathic”, and “sympathetic”. Adjectives were rated on a self-report rating scale, with response options ranging from “0” to “5”. We added scores for the three empathy adjectives together to form an unweighted composite ($M = 6.0$, $SD = 4.03$). Scores on the empathy composite did not differ between men ($M = 5.99$, $SD = 3.91$) and women ($M = 6.02$, $SD = 4.14$; $t(464) = -0.07$, $p = .946$; *Cohen's* $d = 0.01$, 95% CI = [-0.18, 0.19]).

Study B

Participants

Data were collected from 220 participants (*females* = 146) from a mid-sized university in the southeastern United States were included in data analysis. Participants received course credit for participation, and could earn additional compensation based on their performance in the study. We did not collect additional information about diversity for this study. After removing 44 suspicious participants, and an additional 17 participants who had missing data, our final sample included 159 participants in data analysis (age data were not collected).

Procedure

Participants were randomly assigned to a perspective-taking manipulation (imagine other vs. remain objective), as well as a social demand condition (high demand vs. low demand). Participants learned about another (bogus) participant who was suffering from a chronic financial need, and were then provided with an opportunity to share money with them.

Measures

ERQ adjectives. Empathy was measured with six empathy adjectives: “sympathetic”, “compassionate”, “moved”, “softhearted”, “warm”, and “tender”. Participants self-reported their agreement with each item on a Likert scale with response options ranging from 1 (strongly disagree) to 7 (strongly agree). We added scores for the six empathy adjectives together to form an unweighted composite (M = 28.27, SD = 7.20). Women’s (M = 28.78, SD = 7.15) empathy scores were not significantly different than men’s (M = 27.31, SD = 7.24; t(157) = 1.23, p = .22; *Cohen's d* = 0.22, *95% CI* = [-0.11, 0.55]).

Study C

Participants

2,207 participants (*females* = 1,126; M_{age} = 36, SD_{age} = 11.29; due to missing data, age data is based on a slightly smaller sample size of n = 2,197) recruited from Mechanical Turk were included in data analysis. Participants earned up-front payment for participating, and could earn additional compensation based on their performance in the study. We did not collect additional information about diversity for this study. Data from this study were published as part of another manuscript [REDACTED FOR REVIEW].

Procedure

Participants were instructed to watch a short video detailing how hurricanes have negatively impacted people during the 2017 and 2018 hurricane season, and were told that they would later be asked questions about, “people who have been affected by hurricanes, and those who wish to help them.” (see [OSF PAGE REDACTED FOR REVIEW] for more information about the study procedure, including the exact stimuli that participants viewed). Participants reported their experiential emotions immediately after the video.

Measures

ERQ adjectives. Empathy was measured with five empathy adjectives: “compassionate”, “empathic”, “softhearted”, “sympathetic”, and “tender”. Participants were asked to indicate how much they felt each empathy emotion by endorsing each adjective on a Likert scale ranging from 1 (not at all) to 7 (extremely). We added the five items together to form an unweighted composite (M = 27.14, SD = 7.53). Women (M = 28.09, SD = 7.04) scored significantly higher on the composite than men (M = 26.15, SD = 7.9; t(2,205) = 6.11, p < .001; *Cohen's d* = 0.26, *95% CI* = [0.18, 0.34]). Participants were also asked to rate a number of distractor emotion adjective items.

Study D

Participants

1,219 participants were recruited from Mechanical Turk. Participants received up-front payment for participating, and could earn additional compensation based on their performance in the study. We excluded data from *n* = 20 who did not complete the experiment within a predetermined time limit. An additional *n* = 12 were then removed who were either missing gender data or reported that they were non-binary. Finally, we removed *n* = 307 who reported suspicion about the experimental ruse from the state empathy analyses, but included these same suspicious participants in the trait empathy analyses. Please see [OSF PAGE REDACTED FOR REVIEW] for more information regarding exclusion criteria. Due to various missing data from 3 participants, the number of participants included in some analyses was slightly smaller than the overall number of participants.

After subject exclusions, a total of 840 participants were included in state empathy analyses (*females* = 455; *Subject ages: 18 – 24* = 152, *25 – 34* = 391*, 35 – 44* = 165*, 45 – 54* = 84*, 55 – 64* = 34*, 65 – 74* = 11*, Age not provided = 3*), and *N* = 1,147 participants were included in the trait empathy analyses (*females* = 639; *Subject ages: 18 – 24* = 208, *25 – 34* = 529*, 35 – 44* = 236*, 45 – 54* = 118*, 55 – 64* = 46*, 65 – 74* = 19*, Age not provided = 3*).

Procedure

Participants were paired with another (real) subject who happened to be completing the experiment at the same time, and chatted with the other subject while waiting for the experiment to begin. After the experiment began, participants were told they would receive a letter from their interaction partner, and then respond to the letter that they received. Although participants were under the pretense that the note was sent by the person with whom they chatted, all interactions following the initial chat were actually a pre-programmed computer script. Participants were randomly assigned to read a note which described either a medium amount of need, or a high level of need.

Afterwards, participants rated how empathic they felt towards the other subject, given the opportunity to send a note to that person, and to donate money to them. Finally, participants underwent a suspicion probe to determine if participants were aware that the note was staged, or that the experiment was designed to test hypotheses about empathy and prosocial behavior. Participants also completed a number of individual differences measures that are beyond the scope of the present study. See [REDACTED FOR REVIEW] for more details about the experimental procedure, and https://osf.io/76txe/ for the exact stimuli that participants viewed.

Measures

ERQ adjectives. Empathy was measured with five empathy adjectives: "compassionate", "empathic", "sympathetic", "concerned", and "tender”. Participants were asked to indicate how much they felt each empathy emotion by endorsing each adjective on a Likert scale with response options ranging from 1 (strongly disagree) to 7 (strongly agree). We formed an unweighted composite of the five empathy adjectives (*M* = 27.63, *SD* = 5.94). Women (*M* = 28.85, *SD* = 5.48) scored significantly higher on the empathy composite than men (*M* = 26.2, *SD* = 6.15) (*t*(835) = 6.59, *p* < .001; *Cohen's d* = 0.46, *95% CI* = [0.32, 0.60]). Participants were also asked to rate a number of distractor emotion adjective items.

Interpersonal Reactivity Index. Trait empathy was measured with seven items from the empathic concern subscale of the Interpersonal Reactivity Index (Davis, 1980). The seven items were: “I often have tender, concerned feelings for people less fortunate than me”, “Sometimes I don't feel very sorry for other people when they are having problems" (reverse scored), "When I see someone being taken advantage of, I feel kind of protective towards them", "Other people's misfortunes do not usually disturb me a great deal" (reversed), "When I see someone being treated unfairly, I sometimes don't feel very much pity for them" (reversed), "I am often quite touched by things that I see happen", and "I would describe myself as a pretty soft-hearted person." Participants endorsed each item on a 5-point Likert scale with items ranging from 1 (does not describe me at all) to 5 (describes me completely). We formed an unweighted composite of the five empathy adjectives (*M* = 27.6, *SD* = 5.45). Females (*M* = 29.02, *SD* = 4.94) scored significantly higher on the empathy composite than males (*M* = 25.8, *SD* = 5.54) (*t*(1,142) = 10.39, *p* < .001; *Cohen's d* = 0.62, *95% CI* = [0.50, 0.74]). Although we measured all 28 items from the IRI, we did not include the other 21 items in analysis since they measure traits other than empathic concern.

Study E

Participants

Data were collected from 253 participants from a mid-sized university in the Southeastern United States. Participants received course credit for participating. We removed participants who expressed suspicion about the experimental protocol (*n* = 18), or received improper condition assignments due to an error in protocol (*n* = 7). The final sample size included in analyses was 228 (*females* = 154). We did not collect additional information about diversity, nor age, for this study.

Procedure

Participants were informed they were participating in a study that investigated whether simply thinking about moral issues influences subsequent social judgments. Before the experiment began, participants were randomly assigned to either a public condition, in which they completed the experimental tasks through an interview with the experimenter, or a private condition, in which they completed the tasks privately.

Participants were then introduced to two social rules and asked to come up with examples of times that each rule has been relevant to episodes from their personal life and the public sphere. Participants were randomly assigned to one of two experimental conditions that manipulated the social rules that they learned about. In the golden rule conditions, one of the social rules was the golden rule (i.e., the principle of treating others as you want to be treated), whereas in the control condition the social rules pertained to neither the golden rule nor helping people. After being introduced to the social rules, the participant was asked to indicate whether they personally believed that each social rule is good.

Next, the participant read two essays that were apparently written by fellow University of Miami students during a previous study. The participants of that study were apparently told to write about the last time that they felt either helpless or inspired. After reading each essay, the subject self-reported the emotions they felt. Participants were then told that the experiment had concluded, but before departing, the experimenter administered to the subject a letter from the professor in charge of the experiment that is putatively unrelated to the study. The letter contained an appeal to volunteer time on behalf of the author of the essay about feeling helpless (e.g., help with tasks related to fundraising efforts). The participant then indicated on a form whether they were willing to help, and, if so, for how many hours. The participants was told before they made this decision that the experimenter will examine the form and send it to the essay narrator if it indicates a commitment to help. Please see https://osf.io/4cya6/ for full details of the study protocol, and all measures included in analyses.

Measures

ERQ adjectives. Experiential empathy was measured with four empathy adjectives from the ERQ: “empathic”, "compassionate", "sympathetic", and "concerned". Participants were asked to indicate how much they felt each empathy emotion by endorsing each adjective on a 7-point scale ranging from 1 (not at all) to 7 (extremely). The four items were added together to form an unweighted composite (*M* = 22.35, *SD* = 4.4). Females (*M* = 23.07, *SD* = 4.22) scored higher than males (*M* = 20.96, *SD* = 4.44) on an observed composite of the four empathy adjectives (*t*(220) = 3.47, *p* < .001; *Cohen's d* = 0.49, *95% CI* = [0.21, 0.77]). Participants were also asked to rate a number of distractor emotion adjective items.

Study F

Participants

Data was collected from *N* = 814 participants. Participants were awarded $1.00 for completing the study, and could earn up to an additional $2.00 in bonus money. 12 participants were removed from data analysis due to missing data, resulting in a final total of 802 participants (*females* = 405; M_{age} = 36.59, SD_{age} = 11.15; *Asian* = 64, *Black* = 65, *Caucasian* = 619, *Hispanic* = 42, *Native American* = 4, *Other* = 8).

Procedure

After providing consent, participants provided their demographic information, and then completed a number of self-report scales related to altruism, religiosity, their personal strivings, psychopathy, and personality (see [REDACTED FOR REVIEW], for a description of all measures). Participants then watched an Oxfam charity video about the plight of children suffering during war (www.youtube.com/watch?v=6b-jmcZJVEk), were told that they would be quizzed about their impressions of the video afterwards, and were then given the opportunity to donate money to Oxfam.

Measures

ERQ adjectives. Experiential empathy was measured with five empathy adjectives: “Compassionate,” “empathic,” “tender,” “softhearted,” and “sympathetic.” Participants were asked to indicate how much they felt each empathy emotion by endorsing each adjective on a 7-point scale ranging from 1 (not at all) to 7 (extremely). The five items were added together to form an unweighted composite (*M* = 28.17, *SD* = 7.25). Females (*M* = 30.12, *SD* = 6.13) scored higher than males (*M* = 26.18, *SD* = 7.75) on an observed composite of the five empathy adjectives (*t*(800) = 7.99, *p* < .001; *Cohen's d* = 0.56, *95% CI* = [0.42, 0.70]). Participants were also asked to rate a number of distractor emotion adjective items.

Study G

Participants

Data were collected from 168 undergraduate students (age data not collected; *females* = 100) recruited from an online subject pool at a southeastern university in the United States. Participants received course credit for completing the study. We did not collect additional information about diversity for this study. We a priori excluded international students who reported they were from an East Asian country to avoid confounding the broader cross-national project design.

Procedures

Participants completed the study online in Qualtrics in exchange for course credit. Participants were asked to "to "think of a time that a close other person did something to upset you, hurt you, or otherwise commit an offense that caused a rift in your relationship." Afterwards, they completed several scales that measured their forgiveness towards the transgressor, as well as other individual differences measures (please see Studies 4 and 5 from [REDACTED FOR REVIEW] for more details about measures that were collected in this study).

Measures

Interpersonal Reactivity Index. Trait empathy was measured using the seven items from the empathic concern subscale of the IRI. Items on the scale ranged from 1 (doesn't describe me at all) to 5 (describes me very well). We formed a weighted composite by adding the seven items together (*M* = 3.62, *SD* = 0.67). Women (*M* = 3.62, *SD* = 0.65) did not differ from men (*M* = 3.61, *SD* = 0.69) in their trait empathic concern *t*(324) = 0.11, *p* = .911; *Cohen's d* = 0.01, *95% CI* = [-0.21, 0.23]).

Study H

Participants

158 undergraduate students (*females* = 86; age data not collected) were recruited from an online subject pool at a university in western Japan. Participants received both course credit, and financial compensation for completing the study. We did not collect additional information about diversity for this study.

Procedures

The study procedure for Study H was nearly the same as Study G, except that participants completed the study in the laboratory in exchange for 700 JPY. Participants completed the IRI along with a battery of questionnaires that are not included in the current study (please see Studies 4 and 5 from [REDACTED FOR REVIEW] for details about measures that were collected in this study).

Measures

Interpersonal Reactivity Index. Trait empathy was measured using the seven items from the empathic concern subscale of the IRI. Participants were asked to indicate how much they felt each empathy emotion by endorsing each adjective on a 7-point scale ranging from 1 (doesn't describe me at all) to 5 (describes me very well). We formed a weighted composite by adding the seven items together (*M* = 3.42, *SD* = 0.63). Women (*M* = 3.43, *SD* = 0.60) did not differ from men (*M* = 3.41, *SD* = 0.68) in their trait empathic concern *t*(156) = 0.18, *p* = .858; *Cohen's d* = 0.03, *95% CI* = [-0.15, 0.21]).

Study I

Participants

862 Prolific workers were recruited online. Participants received financial compensation for completing the study. We removed *n* = 20 for failing to complete the entire study, and an additional *n* = 29 for failing to provide information about their sex, resulting in a final sample of *n* = 812 (*females* = 403; M_{age} = 37.85, SD_{age}= 13.09; 9 = *American Indian/Alaskan Native*, 77 = *Asian*, 77 = *Black or African American*, 596 = *White*, 29 = *More than one race*, 19 = *Other*, 4 = *Prefer not to answer*).

Procedures

Data collection took place in December 2023. After providing consent, participants were informed that they would complete a study in which they would read about another person, and then report the emotions that they're feeling after reading the story. Participants then read about a story written by a person seeking financial help on HandUp.org, a charitable navigator. After reading the story, participants then self-reported the emotions they were currently feeling via the ERQ; were provided the opportunity to donate money to the person; completed the IRI; and finally, they provided their demographic information and were debriefed.

Measures

ERQ adjectives. In wave one, experiential empathy was measured with five empathy adjectives from the ERQ: "Compassionate", "moved", "tender", "softhearted", and "sympathetic". In wave two, experiential empathy was measured with the same five empathy adjectives, plus three additional adjectives: "Warm", "empathic", and "concerned". Participants were asked to indicate how much they felt each empathy emotion by endorsing each adjective on a 7-point scale ranging from 1 (not at all) to 7 (extremely). Participants were also asked to rate a number of distractor emotion adjective items. We formed a weighted state empathy composite by adding the five items together (*M* = 4.76, *SD* = 1.54).Women (*M* = 4.96, *SD* = 1.54) scored higher than men (*M* = 4.57, *SD* = 1.51) on an observed composite of the five empathy adjectives (*t*(810) = 3.67, *p* < .001; *Cohen's d* = 0.26, *95% CI* = [0.12, 0.40]).

Interpersonal Reactivity Index. Trait empathy was measured using the seven items from the empathic concern subscale of the IRI. Participants were asked to indicate how much they felt each empathy emotion by endorsing each adjective on a 5-point scale ranging from 1 (doesn't describe me at all) to 5 (describes me very well). We formed a weighted composite by adding the seven items together (*M* = 3.95, *SD* = 0.76). Women (*M* = 4.13, *SD* = 0.73) scored significantly higher than men (*M* = 3.78, *SD* = 0.76) in their trait empathic concern *t*(810) = 6.70, *p* < .001; *Cohen's d* = 0.47, *95% CI* = [0.33, 0.61]).

Study J

Participants

Data were collected from *n* = 300 undergraduate students. Participants received course credit for completing the study. We removed data from n = 5 participants who indicated they were non-binary, leaving a final sample of *n* = 295 (*females* = 221; M_{age} = 20.51, SD_{age} = 2.13, *n* = 2 failed to provide age data) recruited from an online subject pool at a southwestern university in the United States. We did not collect additional information about diversity for this study.

Procedures

Participants were seated at a laptop computer for the duration of the study. After providing consent, participants were informed that the ostensive purpose of the study was to source video content for a high school civics course. Participants then watched the first six minutes of an approximately 17-minute-long TED Talk about effective altruism, "The Why and How of Effective Altruism" by Peter Singer (https://www.youtube.com/watch?v=Diuv3XZQXyc); answered 10 multiple-choice quiz questions about the content of the video that they had watched; and completed the Emotion Reaction Questionnaire.

Participants were then told that they could continue watching the TED Talk for as long as they wanted, or simply advance the page and watch none of the remainder of the video. This video and quiz procedure was repeated for a second TED Talk video, "New Video Technology That Reveals an Object's Hidden Properties" by Abe Davis (https://www.youtube.com/watch?v=npNYP2vzaPo), where participants once again had the opportunity to watch as much or as little of the remainder of the video as they wanted after the quiz and Emotion Reaction Questionnaire were complete.

Measures

ERQ adjectives. Experiential empathy was measured with five empathy adjectives from the ERR: "Compassionate", "empathic", "tender", "softhearted", and "sympathetic". Participants were asked to indicate how much they felt each empathy emotion by endorsing each adjective on a 7-point scale ranging from 1 (not at all) to 7 (extremely). The five items were added together to form an unweighted composite (*M* = 4.42, *SD* = 1.40). Females (*M* = 4.50, *SD* = 1.38) scored higher than males (*M* = 4.18, *SD* = 1.47) on an observed composite of the five empathy adjectives (*t*(293) = 1.68, *p* = .094; *Cohen's d* = 0.23, *95% CI* = [-0.03, 0.49]). Participants were also asked to rate a number of distractor emotion adjective items.

Interpersonal Reactivity Index. Trait empathy was measured using the seven items from the empathic concern subscale of the IRI. Participants were asked to indicate how much they felt each empathy emotion by endorsing each adjective on a 5-point scale ranging from 0 (doesn't describe me at all) to 4 (describes me very well). We formed a weighted composite by adding the seven items together (*M* = 3.90, *SD* = 0.65). Women (*M* = 4.01, *SD* = 0.60) scored significantly higher than men (*M* = 3.57, *SD* = 0.70) in their trait empathic concern *t*(293) = 5.31, *p* < .001; *Cohen's d* = 0.71, *95% CI* = [0.44, 0.98]).

Study K

Participants

N = 400 Prolific workers were recruited online. Participants received financial compensation for completing the study. We removed *n* = 10 for failing to provide information about their sex, resulting in a final sample of *n* = 390 (*females* = 215; M_{age} = 35.94, SD_{age} = 12.46; 41 = *Asian*, 41 = *Black or African American*, 269 = *White*, 23 = *More than one race*, 12 = *Other*, 4 = *Prefer not to answer*).

Procedure

The data collection procedure was identical to Study I, except that participants completed an eight-item version of the ERQ in Study K, rather than the five-item version of the ERQ included in Study I.

Measures

ERQ adjectives. In wave one, experiential empathy was measured with five empathy adjectives from the ERQ: "Compassionate", "moved", "tender", "softhearted", "sympathetic", "warm", "empathic", and "concerned". Participants were asked to indicate how much they felt each empathy emotion by endorsing each adjective on a 7-point scale ranging from 1 (not at all) to 7 (extremely). Participants were also asked to rate a number of distractor emotion adjective items. We formed an eight-item composite from responses to the empathic concern items (*M* = 4.58, *SD* = 1.46). Women (*M* = 4.82, *SD* = 1.36) scored higher than men (*M* = 4.29, *SD* = 1.52) on an observed composite of the eight empathy adjectives (*t*(388) = 3.58, *p* < .001; *Cohen's d* = 0.36, *95% CI* = [0.16, 0.56]).

Interpersonal Reactivity Index. Trait empathy was measured using the seven items from the empathic concern subscale of the IRI. Participants were asked to indicate how much they felt each empathy emotion by endorsing each adjective on a 5-point scale ranging from 1 (doesn't describe me at all) to 5 (describes me very well). We formed a weighted composite by adding the seven items together (*M* = 3.95, *SD* = 0.74). Women (*M* = 4.13, *SD* = 0.69) scored significantly higher than men (*M* = 3.73, *SD* = 0.75) in their trait empathic concern *t*(388) = 5.51, *p* < .001, *Cohen's d* = 0.56, *95% CI* = [0.36, 0.76]).

Study results

Study A

Analysis code

Analyses for this study are included in an R script titled "Study_A_MI.R."

Model specification

Experimental manipulations and regressions

We regressed the following covariates on the empathy factor: (1) Whether empathy was manipulated by having the subject read about a victim who was experiencing a break-up (*control* = 0, *empathy manipulation* = 1), (2) The quality of the subject's interaction with the empathic target in a trust game (no trust game vs. fair partner vs. generous partner vs. very generous partner) using *k*-1 dummy codes, with participants who did not play the trust game serving as the reference group, (3) The prospect of future trust game encounters with the victim (*no future encounter* = 0, *future encounter* = 1), (4) Whether or not the victim was a stranger or a friend of the subject (*stranger* = 0, *friend* = 1), (5) The experiment that participants participated in (experiment 2 vs. experiment 3 vs. experiment 4 vs. experiment 5) using *k*-1 dummy codes, with participants in experiment 5 serving as the reference group.

Results

Table S2

Results of AOAA anchor selection analyses for Study A, and the dMACs for the configural model.

	Factor loading	*Satorra-Bentler χ2 (df)*	*p-value*	*dMAC*
Sympathetic*	0.842	3.497 (5)	.624	0.19
Compassionate	0.777	3.00 (5)	.700	0.04
Empathic	0.767	1.864 (5)	.868	0.06

Note: Factor loadings are the standardized loadings from the constrained model. Satorra-Bentler χ2 (df) is the change in χ2 value for the fully constrained model versus a model wherein the parameters for each item are freely estimated. P-value is the p-value associated with the difference between the models. dMACs are the degree of measurement non-equivalence for each item from the configural model with males serving as the reference group. For each study, the non-invariant item with the highest factor loading is marked by an asterisk. Items that showed evidence of non-invariance are marked by a hashtag.

Table S3

AFIs and ΔAFIs for the measurement invariance analyses in Study A.

	df	*Robust CFI*	*Robust RMSEA*	*SRMR*
Unidimensional configural	36	1.000	0.008	0.002
ΔThreshold	9	0.0002	-0.007	0.0
ΔMetric	2	0.0	-0.001	0.0009
ΔScalar	3	0.0	0.0	0.003
ΔStrict	2	0.0	0.0	0.0
Strict	52	1.000	0.0	0.006

Table S4

Six largest residual correlations for the multidimensional configural model in Study A.

Path	*Residual correlation*	*Modification indice ($\Delta\chi^2$)*
Males		
Sympathetic ~~ compassionate	0.005	0.66
Sympathetic ~~ empathic	-0.003	0.34
Compassionate ~~ empathic	-0.002	0.06
Females		
Compassionate ~~ empathic	0.003	0.18
Sympathetic ~~ empathic	-0.001	0.48
Sympathetic ~~ compassionate	-0.001	0.55

Note: Modification indice is the $\Delta\chi^2$ after adding the covariance to the model. ~~ indicates a shared covariance between adjectives. We only report three residual correlations for each group because there were only three indicators, so only three covariances are possible.

Study B

Analysis code

Analyses for this study are included in an R script titled “Study_B_MI.R.”

Model specification

Experimental manipulations and regressions

To control for the effect of the experimental instructions, we regressed the perspective-taking (*remain objective* = 0, *imagine other* = 1) and social demand manipulations (*low demand* = 0, *high demand* = 1) on the latent variables.

Category collapsing

None of the males, and only two of the females (1.9% of all females), endorsed a response category less than “3” for the item “sympathetic,” so responses of “1”, “2”, or “3” for “sympathetic” were collapsed into a single response category of “3” for both men and women.

Results

Table S5

Results of AOAA anchor selection analyses for Study B, and the dMACs for each configural model.

	Factor loading	*Satorra-Bentler* χ*2 (df)*	*p-value*	*dMAC*
Unidimensional model				
Sympathetic	0.571	4.134 (4)	.388	0.37
Compassionate*	0.758	2.679 (6)	.848	0.20
Moved	0.736	1.409 (6)	.965	0.05
Tender	0.664	7.065 (6)	.315	0.13
Softhearted	0.693	3.502 (6)	.744	0.15
Warm	0.444	7.363 (6)	.289	0.23
Multidimensional model				
Sympathetic	0.552	1.08 (4)	.897	0.36
Compassionate*	0.756	2.501 (6)	.868	0.20
Moved	0.731	1.782 (6)	.939	0.05
Tender	0.701	6.331 (6)	.387	0.13
*Softhearted**	0.744	2.883 (6)	.823	0.13
Warm	0.465	4.999 (6)	.544	0.20

Note: Factor loadings are the standardized loadings from the constrained model. Satorra-Bentler χ2 (df) is the change in χ2 value for the fully constrained model versus a model wherein the parameters for each item are freely estimated. P-value is the p-value associated with the difference between the models. dMACs are the degree of measurement non-equivalence for each item from the configural model with males serving as the reference group. The non-invariant item with the highest factor loading is marked by an asterisk. Items that showed evidence of non-invariance are marked by a hashtag. For the multidimensional model, anchor items that loaded onto the sympathy factor are bolded, and items that loaded onto the tenderness factor are in italics.

Table S6

AFIs and ΔAFIs for the measurement invariance analyses in Study B.

	df	*Robust CFI*	*Robust RMSEA*	*SRMR*
Measurement invariance models				
Unidimensional configural*	38	0.997	0.024	0.067
Multidimensional configural	32	1.00	0.00	0.063
ΔThresholds constrained	22	-0.0001	-0.005	0.000
ΔMetric	5	0.002	-0.008	0.004
ΔScalar	5	0.001	-0.012	0.008
ΔStrict	6	-0.010	0.033	0.021
Strict	76	0.990	0.033	0.100

Note: The configural model that was retained for analyses is marked by an asterisk.

Table S7

Ten largest residual correlations for the multidimensional configural model in Study B.

Path	*Residual correlation*	*Modification indice ($\Delta\chi^2$)*
Males		
Sympathetic ~~ warm	-0.17	1.545
Softhearted ~~ warm	0.11	1.032
Tender ~~ moved	-0.09	1.00
Softhearted ~~ compassionate	-0.06	0.620
Tender ~~ softhearted	0.05	0.436
Females		
Sympathetic ~~ warm	-0.26	4.107
Sympathetic ~~ tender	-0.11	1.121
Tender ~~ moved	-0.10	2.064
Tender ~~ warm	0.08	1.489
Sympathetic ~~ compassionate	0.08	2.639

Note: Modification indice is the $\Delta\chi^2$ after adding the covariance to the model. ~~ indicates a shared covariance between adjectives.

Study C

Analysis code

Analyses for this study are included in an R script titled "Study_C_MI.R." Output from the data file (i.e., the R workspace associated with the data) is saved in i

Results

For the multidimensional model, the correlation between the sympathy and tenderness factors was significant (

Table S8

Results of AOAA anchor selection analyses for Study C, and the dMACs for each configural model.

	Factor loading	*Satorra-Bentler χ2 (df)*	*p-value*	*dMAC*
Unidimensional model				
Sympathetic	0.925	5.331 (6)	.502	0.23
Compassionate	0.918	3.885(6)	.692	0.21
Empathic	0.899	4.797 (6)	.570	0.19
Softhearted*	0.926	4.975 (6)	.547	0.19
Tender	0.896	2.247 (6)	.896	0.20
Multidimensional model				
Sympathetic*	0.929	6.049 (6)	.418	0.23
Compassionate	0.922	3.671 (6)	.721	0.21
Empathic	0.902	5.09 (6)	.532	0.19
*Softhearted**	0.936	5.145 (6)	.525	0.19
Tender	0.903	5.497 (6)	.482	0.20

Note: Factor loadings are the standardized loadings from the constrained model. Satorra-Bentler χ2 (df) is the change in χ2 value for the fully constrained model versus a model wherein the parameters for each item are freely estimated. P-value is the p-value associated with the difference between the models. dMACs are the degree of measurement non-equivalence for each item from the configural model with males serving as the reference group. For each study, the non-invariant item with the highest factor loading is marked by an asterisk. Items that showed evidence of non-invariance are marked by a hashtag. For the multidimensional model, anchor items that loaded onto the sympathy factor are bolded, and items that loaded onto the tenderness factor are in italics.

Table S9

AFIs and ΔAFIs for the measurement invariance analyses in Study C.

	df	*Robust CFI*	*Robust RMSEA*	*SRMR*
Unidimensional configural	10	1.00	0.053	0.011
Multidimensional configural*	8	1.00	0.031	0.006
ΔThreshold	16	0.00	-0.014	0.00
ΔMetric	7	0.00	-0.001	0.00
ΔScalar	3	0.00	-0.003	0.00
ΔStrict	5	0.00	-0.0004	0.002
Strict	39	1.00	0.013	0.008

Note: The configural model that was retained for analyses is marked by an asterisk.

Table S10

Ten largest standardized residual covariances for the multidimensional configural model in Study C.

Path	*Standardized residual covariance*	*Modification indice ($\Delta\chi^2$)*
Males		
Sympathetic ~~ compassionate	-7.08	3.44
Compassionate ~~ empathic	6.50	3.68
Tender ~~ sympathetic	3.97	1.55
Tender ~~ empathic	-3.85	1.39
Softhearted ~~ empathic	-3.82	0.73

Females		
Sympathetic ~~ compassionate	-7.57	4.57
Softhearted ~~ empathic	-5.67	1.45
Tender ~~ empathic	-4.31	1.35
Compassionate ~~ empathic	4.25	1.92
Softhearted ~~ sympathetic	2.93	0.54

Note: Modification indice is the $\Delta\chi^2$ after adding the covariance to the model. ~~ indicates a shared covariance between adjectives.

Permutation test

For the failed configural model, we conducted a permutation test to determine whether the model failure was due to a group discrepancy, or an overall approximation discrepancy. During permutation testing, we occasionally encountered the unexpected issue of sparse category endorsement in the permuted datasets. Since permutation tests rely on repeatedly shuffling group assignment to obtain the null chi-squared distribution, items which have few endorsements for extreme response categories may be randomly reshuffled such that one of the groups has endorsements for the extreme response category, while the other group has no responses for the extreme response category. Because some of the items in our data had few endorsements for higher (lower) response categories, some of the permuted datasets produced sparse categories which caused the permutation test to terminate prematurely. In the event of sparse categories, Kite, Jorgensen, and Chen (2018) recommend discarding the permuted dataset with sparse categories, and simply re-permuting the dataset again (see Kite et al., 2018, Appendix A, for an elaboration on this solution and a discussion of other potential solutions). The permuteMeasEq() function includes the argument maxSparse to specify the maximum number of times that the dataset can re-permuted when at least one response category is unobserved in at least one group.

We ran a permutation test by generating 1,000 randomly permuted datasets and setting maxSparse = 20[1]. The results of the permutation test indicated that there was no evidence against the null of group discrepancy (*Scaled* $\chi^2(8) = 63.510$, $p = 0.995$), suggesting that the discrepancy was due to an overall approximation discrepancy in the baseline model, rather than a configural misspecification for men versus women.

Study D-1

Analysis code

Analyses for this study are included in an R script titled "Study_D-1_MI.R."

Model specification

Experimental manipulations and regressions

We regressed the empathy factor on (1) The instructions condition to which each subject was assigned (self vs. other vs. reman objective vs. control) using *k*-1 dummy codes, with

[1] MaxSparse = 20 was itself a data-driven decision, as we initially attempted to use the default setting (MaxSparse = 10), but this resulted in model failure.

participants in the control condition serving as the reference group, and (2) The need state of the empathic targets (*low need* = 0, *high need* = 1).

Results

Table S11

Results of AOAA anchor selection analyses and dMACs for the configural models from Study D-1.

	Factor loading	*Satorra-Bentler χ2 (df)*	*p-value*	*dMAC*
Unidimensional model				
Sympathetic	0.877	3.18 (6)	.786	0.42
Compassionate*	0.888	5.56 (6)	.474	0.78
Empathic	0.730	4.98 (6)	.547	0.49
Tender	0.668	1.46 (6)	.962	0.40
Concerned	0.625	5.22 (6)	.516	0.24
Multidimensional model				
Sympathetic	0.878	6.49 (6)	.371	0.43
Compassionate*	0.892	6.01 (6)	.422	0.78
Empathic	0.734	4.81 (6)	.569	0.49
*Tender**	0.700	4.07 (6)	.668	0.56
Concerned	0.657	3.99 (6)	.678	0.84

Note: Factor loadings are the standardized loadings from the constrained model. Satorra-Bentler χ2 (df) is the change in χ2 value for the fully constrained model versus a model wherein the parameters for each item are freely estimated. P-value is the p-value associated with the difference between the models. dMACs are the degree of measurement non-equivalence for each item from the configural model with males serving as the reference group. For each study, the non-invariant item with the highest factor loading is marked by an asterisk. Items that showed evidence of non-invariance are marked by a hashtag. For the multidimensional model, anchor items that loaded onto the sympathy factor are bolded, and items that loaded onto the tenderness factor are in italics.

Permutation test

Which configural null hypothesis is violated for the unidimensional model?

We ran a permutation test of the configural model, generating 1,000 randomly permuted datasets, and setting maxSparse = 20. Test results provided evidence against the null hypothesis that the population model configuration is equally misspecified for both men and women (*Scaled* $\chi^2(32) = 171.099$, $p < .001$), indicating that the χ^2 test was due to a group discrepancy: the configural structure of empathy is different for men and women. Table S10 includes the residual covariances and modification indices for each group.

Table S12

Ten largest residual correlations and modifications indices for the multidimensional configural model in Study D-1.

Path	*Residual correlation*	*Modification indice*
Males		

Empathic ~~ concerned	-0.034	1.045
Sympathetic ~~ tender	0.024	0.824
Compassionate ~~ empathic	0.018	1.368
Sympathetic ~~ concerned	0.011	0.155
Compassionate ~~ tender	-0.009	0.151
Females		
Compassionate ~~ tender	0.058	9.199
Empathic ~~ tender	-0.058	3.017
Sympathetic ~~ empathic	0.045	5.411
Empathic ~~ concerned	-0.027	0.750
Sympathetic ~~ tender	-0.021	0.930

Note: Residuals are Bentler-corrected correlations that were obtained by dividing the elements of both the observed and model implied covariance matrices by the square roots of the corresponding variances of the observed covariance matrix.

Table S13

AFIs and ΔAFIs for the measurement invariance analyses in Study D-1.

	df	*Robust CFI*	*Robust RMSEA*	*SRMR*
Unidimensional configural	42	0.993	0.060	0.028
Multidimensional configural*	32	0.994	0.064	0.020
ΔThreshold	20	0.00	-0.014	0.00
ΔMetric	3	-0.0001	-0.001	0.0002
ΔScalar	3	0.0003	-0.002	0.00
ΔStrict	5	0.0002	-0.003	0.001
Strict	63	0.994	0.044	0.021

Note: The configural model that was retained for analyses is marked by an asterisk.

Study D-2

Analysis code

Analyses for this study are included in an R script titled "Study_D-2_MI.R."

Model specification

Residual correlations between negatively worded items

We included residual correlations between the three negatively worded items for both the unidimensional and sympathy-tenderness models: "Other people's misfortunes do not usually disturb me a great deal," "Sometimes I don't feel very sorry for other people when they are having problems," and "When I see someone being treated unfairly, I sometimes don't feel very much pity for them." We did not regress any predictors on the trait empathy factor(s), as the empathy manipulations should not influence participants' trait empathy scores.

Results

Table S14
Results of AOAA anchor selection analyses for Study D-2, and the dMACs for each configural model.

	Factor loading	*Satorra-Bentler χ2 (df)*	*p-value*	*dMAC*
Unidimensional model				
1. Sometimes I don't feel very sorry for other people when they are having problems.	0.571	8.914 (4)	.063	0.148
2. When I see someone being taken advantage of, I feel kind of protective towards them.	0.767	2.657 (4)	.617	0.345
3. Other people's misfortunes do not usually disturb me a great deal.	0.642	5.480 (4)	.241	0.427
4. When I see someone being treated unfairly, I sometimes don't feel very much pity for them. #	0.704	10.974 (4)	.027*	0.319
5. I often have tender, concerned feelings for people less fortunate than me. *	0.883	4.476 (4)	.345	0.513
6. I am often quite touched by things that I see happen.	0.781	3.428 (4)	.489	0.373
7. I would describe myself as a pretty soft-hearted person.	0.756	4.515 (4)	.341	0.443
Multidimensional model				
1. Sometimes I don't feel very sorry for other people when they are having problems.	0.579	9.173 (4)	.057	0.151
2. When I see someone being taken advantage of, I feel kind of protective towards them. *	0.774	2.608 (4)	.625	0.346
3. Other people's misfortunes do not usually disturb me a great deal.	0.651	6.281 (4)	.179	0.431
4. When I see someone being treated unfairly, I sometimes don't feel very much pity for them #	0.713	14.475 (4)	.006**	0.324
*5. I often have tender, concerned feelings for people less fortunate than me. **	0.890	3.613 (4)	.461	0.515
6. I am often quite touched by things that I see happen.	0.785	5.047 (4)	.282	0.374
7. I would describe myself as a pretty soft-hearted person.	0.759	6.733 (4)	.151	0.444

Note: Factor loadings are the standardized loadings from the constrained model. Satorra-Bentler χ2 (df) is the change in χ2 value for the fully constrained model versus a model wherein the parameters for each item are freely estimated. P-value is the p-value associated with the difference between the models. * = *p*-values < .05 ** = *p*-values < .01. dMACs are the degree of measurement non-equivalence for each item from the configural model with males serving as the reference group. For each study, the non-invariant item with the highest factor loading is marked by an asterisk (*). Items that showed evidence of non-invariance are marked by a hashtag (#). For the multidimensional model, anchor items that loaded onto the sympathy factor are bolded, and items that loaded onto the tenderness factor are in italics.

Permutation test

Which configural null hypothesis is violated for the unidimensional model?

We ran a permutation test by generating 1,000 randomly permuted datasets and setting maxSparse = 20. The results of the permutation test indicated that there was no evidence against the null of a group discrepancy (*Scaled* $\chi^2(8) = 151.932$, $p = .781$), suggesting that the discrepancy was due to an overall approximation discrepancy in the baseline model, and that to obtain the true model specification, modifications should be made for the configural model.

Table S15

AFIs and ΔAFIs for the measurement invariance analyses in Study D-2.

	df	*Robust CFI*	*Robust RMSEA*	*SRMR*
Unidimensional configural*	22	0.997	0.061	0.029
Multidimensional configural	20	0.997	0.063	0.028
ΔThreshold	14	-0.0001	-0.012	0.00
ΔMetric	6	-0.0003	-0.0017	0.0011
ΔScalar	6	-0.0001	-0.0022	0.0005
ΔStrict#	6	0.0002	-0.0027	0.0011
ΔPartial Strict#	4	-0.002	0.008	0.005
Scalar	48	0.997	0.045	0.031

Note: Models that failed invariance tests are marked by a hashtag. In the event that a partial invariant model is reported, the fit statistics represent the difference between the partial model, and the last non-invariant model.

Table S16

Ten largest standardized residual covariances for the failed configural model in Study D-2.

Path	*Standardized residual covariance*	*Modification indice ($\Delta\chi^2$)*
Males		
1. “I am often quite touched by things that I see happen” ~~ “I often have tender, concerned feelings for people less fortunate than me”	7.844	1.495
2. “When I see someone being treated unfairly, I sometimes don't feel very much pity for them” ~~ “I would describe myself as a pretty soft-hearted person”	-7.658	6.857
3. “When I see someone being taken advantage of, I feel kind of protective towards them” ~~ “When I see someone being treated unfairly, I sometimes don't feel very much pity for them”	6.387	7.114
4. “I often have tender, concerned feelings for people less fortunate than me” ~~ “Other people's misfortunes do not usually disturb me a great deal”	5.707	5.396
5. “I often have tender, concerned feelings for people less fortunate than me” ~~ “I would describe myself as a pretty soft-hearted person”	-4.598	2.987

Females		
1. “When I see someone being taken advantage of, I feel kind of protective towards them” ~~ “When I see someone being treated unfairly, I sometimes don't feel very much pity for them”	6.840	10.774
2. “I am often quite touched by things that I see happen” ~~ “I would describe myself as a pretty soft-hearted person”	5.445	8.752
3. “When I see someone being treated unfairly, I sometimes don't feel very much pity for them” ~~ “I would describe myself as a pretty soft-hearted person”	4.742	3.240
4. “When I see someone being taken advantage of, I feel kind of protective towards them” ~~ “I would describe myself as a pretty soft-hearted person”	-4.501	5.036
5. “Sometimes I don't feel very sorry for other people when they are having problems” ~~ “I would describe myself as a pretty soft-hearted person”	4.411	4.244

Note: Standardized residual covariance is the *Z*-transformed residual covariance. Modification indice is the $\Delta\chi^2$ after adding the covariance to the model.

Study E

Analysis code

Analyses for this study are included in an R script titled “Study_E_MI.R.”

Model specification

Experimental manipulations and regressors

We regressed the empathy factor on two exogenous predictors: participants’ assignment to either the golden rule or tolerance condition (*tolerance* = 0, *golden rule* = 1), and their assignment to either the public or private post-interview questionnaire (*private* = 0, *public* = 1).

Category collapsing

For the adjective “sympathetic”, two female participants (1.35% of females) and one male subject (2.5% of males) endorsed a “1”. For the same item, zero female participants and one male subject endorsed a “2” (2.5% of males). We recoded all responses less than “2” as “2” for “sympathetic” in both groups.

Results

Table S17

Results of AOAA anchor selection analyses for Study E, and the dMACs for each configural model.

	Factor loading	*Satorra-Bentler χ2 (df)*	*p-value*	*dMAC*
Sympathetic*	0.667	4.154 (5)	.528	0.515
Compassionate	0.616	11.181 (6)	.083	0.039
Empathic	0.556	5.503 (6)	.481	0.066
Concerned#	0.650	14.259 (5)	.014*	0.360

Note: Factor loadings are the standardized loadings from the constrained model. Satorra-Bentler χ2 (df) is the change in χ2 value for the fully constrained model versus a model wherein the parameters for each item are freely estimated. P-value is the p-value associated with the difference between the models. * = *p*-values < .05. dMACs are the degree of measurement non-equivalence for each item from the configural model with males serving as the reference group. For each study, the non-invariant item with the highest factor loading is marked by an asterisk. Items that showed evidence of non-invariance are marked by a hashtag.

Table S18

AFIs and ΔAFIs for the measurement invariance analyses in Study E.

	df	*Robust CFI*	*Robust RMSEA*	*SRMR*
Unidimensional configural	16	0.998	0.019	0.020
ΔThreshold	14	-0.0004	-0.003	0.0
ΔMetric	3	0.002	-0.015	0.005
ΔScalar	3	-0.02	0.04	0.02
ΔStrict	4	-0.015	0.01	-0.01
Strict	40	0.965	0.052	0.034

Table S19

Ten largest residual correlations for the multidimensional configural model in Study E.

Path	*Residual correlation*	*Modification indice ($\Delta\chi^2$)*
Males		
Sympathetic ~~ concerned	-0.02	0.554
Sympathetic ~~ compassionate	0.01	0.115
Concerned ~~ empathic	0.01	0.196
Empathic ~~ sympathetic	-0.01	1.110
Concerned ~~ compassionate	-0.004	0.576
Females		
Sympathetic ~~ concerned	-0.06	0.081
Empathic ~~ sympathetic	0.03	0.008
Concerned ~~ compassionate	0.03	0.017
Concerned ~~ empathic	-0.02	0.048
Empathic ~~ compassionate	-0.01	0.015

Note: Modification indice is the $\Delta\chi^2$ after adding the covariance to the model. ~~ indicates a shared covariance between adjectives.

Study F

Analysis code

Analyses for this study are included in an R script titled "Study_F_MI.R."

Model specification

Experimental manipulations and regressors

This study did not include any experimental manipulations that needed to be controlled for in analyses.

Results

Table S20

Results of AOAA anchor selection analyses for Study F, and the dMACs for each configural model.

	Factor loading	*Satorra-Bentler χ2 (df)*	*p-value*	*dMAC*
Unidimensional model				
Sympathetic	0.939	8.715 (6)	.190	0.48
Compassionate*	0.941	1.872 (6)	.931	0.52
Empathic	0.847	1.072 (6)	.983	0.40
Tender	0.909	5.235 (6)	.514	0.47
Softhearted	0.923	0.623 (6)	.996	0.48
Multidimensional model				
Sympathetic*	0.952	8.061 (6)	.234	0.48
Compassionate	0.952	2.323 (6)	.888	0.52
Empathic	0.857	1.910 (6)	.928	0.40
Tender	0.927	6.409 (6)	.379	0.48
*Softhearted**	0.950	2.986 (6)	.811	0.48

Note: Factor loadings are the standardized loadings from the constrained model. Satorra-Bentler χ2 (df) is the change in χ2 value for the fully constrained model versus a model wherein the parameters for each item are freely estimated. P-value is the p-value associated with the difference between the models. dMACs are the degree of measurement non-equivalence for each item from the configural model with males serving as the reference group. For each study, the non-invariant item with the highest factor loading is marked by an asterisk. Items that showed evidence of non-invariance are marked by a hashtag. For the multidimensional model, anchor items that loaded onto the sympathy factor are bolded, and items that loaded onto the tenderness factor are in italics.

Table S21

AFIs and ΔAFIs for the measurement invariance analyses in Study F.

	df	*Robust CFI*	*Robust RMSEA*	*SRMR*
Unidimensional configural	10	0.999	0.107	0.025
Multidimensional configural*	8	1.00	0.043	0.011
ΔThreshold	20	0.0	-0.025	0.0
ΔMetric	3	0.0	-0.002	0.0002
ΔScalar	3	0.0	-0.003	0.0002
ΔStrict	5	0.0	0.004	0.005

Strict	39	1.00	0.019	0.016

Note: The configural model that was retained for analyses is marked by an asterisk.

Table S22

Ten largest standardized residual covariances for the failed configural model in Study F.

Path	*Residual correlation*	*Modification indice ($\Delta\chi^2$)*
Males		
Tender ~~ Sympathetic	-0.021	0.092
Tender ~~ Empathic	0.018	1.337
Empathic ~~ Softhearted	-0.013	2.083
Softhearted ~~ Sympathetic	0.01	1.298
Empathic ~~ Compassionate	-0.01	1.967
Females		
Softhearted ~~ Empathic	-0.032	0.626
Tender ~~ Empathic	-0.026	0.237
Empathic ~~ Compassionate	0.018	0.276
Softhearted ~~ Sympathetic	0.012	0.828
Compassionate ~~ Sympathetic	-0.011	0.237

Note: Modification indice is the $\Delta\chi^2$ after adding the covariance to the model.

Permutation test

Which configural null hypothesis is violated for the unidimensional model?

We ran a permutation test by generating 1,000 randomly permuted datasets and setting maxSparse = 20. The results of the permutation test indicated that there was no evidence against the null of group discrepancy (*Scaled* χ^2(x) = 39.284, p = .128), suggesting that the discrepancy was due to an overall approximation discrepancy in the baseline model, and that to obtain the true model specification, modifications should be made for the configural model.

Study G

Analysis code

Analyses for this study are included in an R script titled "Study_G_MI.R."

Model specification

Residual correlations

We included residual correlations between the three negatively worded items for both the unidimensional and sympathy-tenderness models: "Other people's misfortunes do not usually disturb me a great deal," "Sometimes I don't feel very sorry for other people when they are having problems," and "When I see someone being treated unfairly, I sometimes don't feel very much pity for them."

Results

Table S23

Results of AOAA anchor selection analyses for Study G, and the dMACs for each configural model.

	Factor loading	*Satorra-Bentler $\chi2$ (df)*	*p-value*	*dMAC*
Unidimensional model				
Sometimes I don't feel very sorry for other people when they are having problems.	0.658	6.898 (4)	.141	0.144
When I see someone being taken advantage of, I feel kind of protective towards them.	0.625	6.443 (4)	.168	0.057
Other people's misfortunes do not usually disturb me a great deal.	0.681	5.748 (4)	.219	0.139
When I see someone being treated unfairly, I sometimes don't feel very much pity for them.	0.628	0.416 (4)	.981	0.094
I often have tender, concerned feelings for people less fortunate than me.*	0.712	7.461 (4)	.113	0.131
I am often quite touched by things that I see happen.	0.480	5.052 (4)	.282	0.147
I would describe myself as a pretty soft-hearted person.	0.667	4.908 (4)	.297	0.097
Multidimensional model				
Sometimes I don't feel very sorry for other people when they are having problems.	0.672	4.930 (4)	.295	0.131
When I see someone being taken advantage of, I feel kind of protective towards them.	0.610	5.002 (4)	.287	0.056
Other people's misfortunes do not usually disturb me a great deal.*	0.684	4.780 (4)	.311	0.148
When I see someone being treated unfairly, I sometimes don't feel very much pity for them. #	0.635	9.961 (4)	.041*	0.065
*I often have tender, concerned feelings for people less fortunate than me.**	0.728	0.467 (4)	.977	0.133

I am often quite touched by things that I see happen.	0.489	4.690 (4)	.321	0.175
I would describe myself as a pretty soft-hearted person.	0.678	8.526 (4)	.074	0.070

Note: Factor loadings are the standardized loadings from the constrained model. Satorra-Bentler $\chi2$ (df) is the change in $\chi2$ value for the fully constrained model versus a model wherein the parameters for each item are freely estimated. P-value is the p-value associated with the difference between the models. * = *p*-values < .05. dMACs are the degree of measurement non-equivalence for each item from the configural model with males serving as the reference group. For each study, the non-invariant item with the highest factor loading is marked by an asterisk. Items that showed evidence of non-invariance are marked by a hashtag. For the multidimensional model, anchor items that loaded onto the sympathy factor are bolded, and items that loaded onto the tenderness factor are in italics.

Permutation test

Which configural null hypothesis is violated for the multidimensional model?

We ran a permutation test by generating 1,000 randomly permuted datasets and setting maxSparse = 20. The results of the permutation test indicated that there was no evidence against the null of group discrepancy (*Scaled* $\chi^2(20) = 41.896$, $p = .216$), suggesting that the discrepancy was due to an overall approximation discrepancy in the baseline model, and that to obtain the true model specification, modifications should be made for the configural model.

Table S24

AFIs and ΔAFIs for the measurement invariance analyses in Study G.

	df	*Robust CFI*	*Robust RMSEA*	*SRMR*
Unidimensional configural invariance model	22	0.990	0.072	0.050
Multidimensional configural invariance model*	20	0.994	0.059	0.043
ΔThreshold	14	-0.001	-0.008	0.0
ΔMetric	5	0.002	-0.009	0.002
ΔScalar#	5	-0.004	0.009	-0.001
ΔPartial scalar	4	0.001	-0.005	0.001
ΔStrict	7	0.0	-0.003	0.0
Strict	50	0.995	0.034	0.046

Note: The configural model that was retained for analyses is marked by an asterisk. Models that failed invariance tests are marked by a hashtag. In the event that a partial invariant model is reported, the fit statistics represent the difference between the partial model, and the last non-invariant model.

Table S25

Ten largest standardized residual covariances for the failed configural model in Study G.

Path	*Residual correlation*	*Modification indice ($\Delta\chi^2$)*
Males		

“I am often quite touched by things that I see happen” ~~ “Other people's misfortunes do not usually disturb me a great deal”	-0.091	1.499
“When I see someone being treated unfairly, I sometimes don't feel very much pity for them” ~~ “When I see someone being taken advantage of, I feel kind of protective towards them”	0.071	2.222
“I am often quite touched by things that I see happen” ~~ “When I see someone being taken advantage of, I feel kind of protective towards them”	0.071	1.509
“When I see someone being taken advantage of, I feel kind of protective towards them” ~~ “I would describe myself as a pretty soft-hearted person”	-0.060	1.066
“Other people's misfortunes do not usually disturb me a great deal” ~~ “When I see someone being taken advantage of, I feel kind of protective towards them”	-0.056	1.376
Females		
“I am often quite touched by things that I see happen” ~~ “When I see someone being taken advantage of, I feel kind of protective towards them”	-0.134	3.695
“I would describe myself as a pretty soft-hearted person” ~~ “When I see someone being taken advantage of, I feel kind of protective towards them”	-0.109	3.742
“When I see someone being taken advantage of, I feel kind of protective towards them”~~ “I often have tender, concerned feelings for people less fortunate than me”	0.100	10.107
“I would describe myself as a pretty soft-hearted person” ~~ “I am often quite touched by things that I see happen”	0.080	3.504
“When I see someone being taken advantage of, I feel kind of protective towards them”~~ “Sometimes I don't feel very sorry for other people when they are having problems”	-0.061	1.446

Note: Modification indice is the $\Delta\chi^2$ after adding the covariance to the model.

Study H

Analysis code

Analyses for this study are included in an R script titled “Study_H_MI.R.”

Model specification

Residual correlations

We included residual correlations between the three negatively worded items for both the unidimensional and multidimensional models: “Other people's misfortunes do not usually disturb me a great deal,” “Sometimes I don't feel very sorry for other people when they are having problems,” and “When I see someone being treated unfairly, I sometimes don't feel very much pity for them.”

Results

Table S26

Results of AOAA anchor selection analyses for Study H, and the dMACs for each configural model.

	Factor loading	*Satorra-Bentler $\chi 2$ (df)*	*p-value*	*dMAC*
Unidimensional model				
Sometimes I don't feel very sorry for other people when they are having problems.	0.623	5.335 (4)	.255	0.132
When I see someone being taken advantage of, I feel kind of protective towards them. #	0.714	9.874 (4)	.043*	0.116
Other people's misfortunes do not usually disturb me a great deal. *	0.682	3.598 (3)	.308	0.120
When I see someone being treated unfairly, I sometimes don't feel very much pity for them.	0.444	3.660 (4)	.454	0.050
I often have tender, concerned feelings for people less fortunate than me.	0.554	3.577 (4)	.466	0.171
I am often quite touched by things that I see happen.	0.232	3.879 (3)	.275	0.103
I would describe myself as a pretty soft-hearted person.	0.500	3.976 (4)	.409	0.142
Sympathy-tenderness model				
Sometimes I don't feel very sorry for other people when they are having problems.	0.719	4.538 (4)	.338	0.131

When I see someone being taken advantage of, I feel kind of protective towards them.	0.728	4.325 (4)	.364	0.115
Other people's misfortunes do not usually disturb me a great deal. *	0.763	6.956 (3)	.073	0.120
When I see someone being treated unfairly, I sometimes don't feel very much pity for them.	0.485	3.539 (4)	.472	0.051
*I often have tender, concerned feelings for people less fortunate than me. **	0.653	1.355 (4)	.852	0.089
I am often quite touched by things that I see happen.	0.264	1.80 (3)	.615	0.085
I would describe myself as a pretty soft-hearted person.	0.567	1.982 (4)	.739	0.068

Note: Factor loadings are the standardized loadings from the constrained model. Satorra-Bentler χ2 (df) is the change in χ2 value for the fully constrained model versus a model wherein the parameters for each item are freely estimated. P-value is the p-value associated with the difference between the models. * = *p*-values < .05. dMACs are the degree of measurement non-equivalence for each item from the configural model with males serving as the reference group. For each study, the non-invariant item with the highest factor loading is marked by an asterisk. Items that showed evidence of non-invariance are marked by a hashtag. For the multidimensional model, anchor items that loaded onto the sympathy factor are bolded, and items that loaded onto the tenderness factor are in italics.

Table S27

AFIs and ΔAFIs for the measurement invariance analyses in Study H.

	df	*Robust CFI*	*Robust RMSEA*	*SRMR*
Unidimensional configural invariance model	22	0.998	0.033	0.057
Multidimensional configural invariance model*	20	1.00	0.0	0.049
ΔThresholds	12	-0.0002	0.008	0.00
ΔMetric	5	0.0002	-0.008	0.0011
ΔScalar	5	0.0	0.0	0.0004
ΔStrict	7	-0.002	0.018	0.002
Strict	49	0.998	0.018	0.052

Table S28

Ten largest residual correlations for the multidimensional configural model in Study H.

Path	*Residual correlation*	*Modification indice ($\Delta\chi^2$)*
Males		
“I am often quite touched by things that I see happen.” ~~ “Other people's misfortunes do not usually disturb me a great deal.”	-0.13	1.342
“When I see someone being taken advantage of, I feel kind of protective towards them.” ~~ “I am often quite touched by things that I see happen.”	0.11	2.063
“When I see someone being taken advantage of, I feel kind of protective towards them.” ~~ “Other people's misfortunes do not usually disturb me a great deal.”	-0.08	0.508
“When I see someone being taken advantage of, I feel kind of protective towards them. ” ~~ “When I see someone being treated unfairly, I sometimes don't feel very much pity for them.”	0.07	1.467
“I am often quite touched by things that I see happen. ” ~~ “When I see someone being treated unfairly, I sometimes don't feel very much pity for them.”	-0.06	0.206
Females		
“I am often quite touched by things that I see happen.” ~~ “When I see someone being taken advantage of, I feel kind of protective towards them.”	-0.17	1.416

“I often have tender, concerned feelings for people less fortunate than me." ~~ “I would describe myself as a pretty soft-hearted person.”	0.15	0.444
“I often have tender, concerned feelings for people less fortunate than me." ~~ “Other people's misfortunes do not usually disturb me a great deal.”	0.13	0.728
“I am often quite touched by things that I see happen." ~~ “Other people's misfortunes do not usually disturb me a great deal.”	0.11	1.047
“When I see someone being taken advantage of, I feel kind of protective towards them." ~~ “I would describe myself as a pretty soft-hearted person.”	-0.09	0.054

Note: Modification indice is the $\Delta\chi^2$ after adding the covariance to the model. ~~ indicates a shared covariance between adjectives.

Study I-1

Analysis code

Analyses for this study are included in an R script titled “Study_I_MI.R.”

Model specification

Experimental manipulations and regressors

This study did not include any experimental manipulations that needed to be controlled for in analyses.

Results

Table S29

Results of AOAA anchor selection analyses for Study I-1, and the dMACs for each configural model.

	Factor loading	*Satorra-Bentler χ2 (df)*	*p-value*	*dMAC*
Unidimensional model				
Sympathetic*	0.929	10.391 (6)	.109	0.190
Compassionate#	0.888	14.922 (6)	.021*	0.243
Moved	0.836	6.551 (6)	.364	0.134
Softhearted#	0.949	18.203 (6)	.006**	0.171
Tender	0.894	2.012 (6)	.919	0.185
Multidimensional model				
Sympathetic*	0.940	10.598 (6)	.102	0.192
Compassionate#	0.893	15.023 (6)	.020*	0.243
Moved	0.845	9.577 (6)	.144	0.135
Softhearted#	0.965	16.908 (6)	.010*	0.173
*Tender**	0.899	3.929 (6)	.686	0.185

Note: Factor loadings are the standardized loadings from the constrained model. Satorra-Bentler χ2 (df) is the change in χ2 value for the fully constrained model versus a model wherein the parameters for each item are freely estimated. P-value is the p-value associated with the difference between the models. dMACs are the degree of measurement non-equivalence for each item from the configural model with males serving as the reference group. For each study, the non-invariant item with the highest factor loading is marked by an asterisk. Items that showed evidence of non-invariance are marked by a hashtag. For the multidimensional model, anchor items that loaded onto the sympathy factor are bolded, and items that loaded onto the tenderness factor are in italics.

Table S30

AFIs and ΔAFIs for the measurement invariance analyses in Study I-1.

	df	*Robust CFI*	*Robust RMSEA*	*SRMR*
Unidimensional configural	10	0.997	0.095	0.019
Multidimensional configural	8	0.998	0.084	0.016
ΔThreshold	20	-0.0001	-0.031	0.0
ΔPartial threshold	18	0.0	-0.34	0.0
ΔMetric	3	0.0	-0.002	0.0002
ΔScalar	3	0.0	0.0003	0.0
ΔStrict	5	-0.0003	0.007	0.001
ΔPartial strict	3	-0.0003	0.009	0.001
Partial strict	35	0.999	0.057	0.018

Note: The configural model that was retained for analyses is marked by an asterisk. Non-invariant models are marked by a hashtag. In the event that a partial invariant model is reported, the fit statistics represent the difference between the partial model, and the last non-invariant model.

Table S31

Ten largest residual covariances for the multidimensional configural model in Study I-1 .

Path	*Residual covariance*	*Modification indice ($\Delta\chi^2$)*
Males		
Sympathetic ~~ Moved	-0.034	8.415

Tender ~~ Moved	0.024	0.844
Softhearted ~~ Compassionate	-0.019	1.285
Sympathetic ~~ Compassionate	0.015	14.148
Tender ~~ Compassionate	-0.013	2.72
Females		
Sympathetic ~~ Moved	-0.044	6.691
Sympathetic ~~ Compassionate	0.023	3.402
Tender ~~ Compassionate	-0.027	0.727
Softhearted ~~ Moved	0.027	0.002
Moved ~~ Compassionate	-0.025	0.331

Note: Modification indice is the $\Delta\chi^2$ after adding the covariance to the model. ~~ indicates a shared covariance between adjectives.

Permutation test

We ran a permutation test by generating 1,000 randomly permuted datasets and setting maxSparse = 20. The results of the permutation test indicated that there was no evidence against the null of group discrepancy (*Scaled* $\chi^2(8) = 103.367$, $p = .332$), suggesting that the discrepancy was due to an overall approximation discrepancy in the baseline model, rather than a configural misspecification for men versus women.

Study I-2

Analysis code

Analyses for this study are included in an R script titled “Study_I_MI.R.”

Model specification

Residual correlations

We included residual correlations between the three negatively worded items for both the one- and multidimensional models: “Other people's misfortunes do not usually disturb me a great deal,” “Sometimes I don't feel very sorry for other people when they are having problems,” and “When I see someone being treated unfairly, I sometimes don't feel very much pity for them.”

Experimental manipulations and regressions

This study did not include any experimental manipulations that needed to be controlled for in analyses.

Results

Table S32

Results of AOAA anchor selection analyses for Study I-2, and the dMACs for each configural model.

	Factor loading	*Satorra-Bentler $\chi 2$ (df)*	*p-value*	*dMAC*
Unidimensional model				

Sometimes I don't feel very sorry for other people when they are having problems.	0.593	5.813 (4)	.214	0.245
When I see someone being taken advantage of, I feel kind of protective towards them.	0.795	2.891 (4)	.576	0.258
Other people's misfortunes do not usually disturb me a great deal.	0.752	4.225 (4)	.376	0.362
When I see someone being treated unfairly, I sometimes don't feel very much pity for them.	0.513	5.851 (4)	.211	0.142
I often have tender, concerned feelings for people less fortunate than me. *	0.908	7.706 (4)	.103	0.308
I am often quite touched by things that I see happen.	0.874	6.087 (4)	.193	0.370
I would describe myself as a pretty soft-hearted person.	0.821	7.324 (4)	.120	0.304
Sympathy-tenderness model				
Sometimes I don't feel very sorry for other people when they are having problems.	0.589	6.094 (4)	.192	0.227
When I see someone being taken advantage of, I feel kind of protective towards them. *	0.785	3.612 (4)	.461	0.246
Other people's misfortunes do not usually disturb me a great deal.	0.743	7.251 (4)	.123	0.345
When I see someone being treated unfairly, I sometimes don't feel very much pity for them.	0.511	2.428 (4)	.658	0.122

*I often have tender, concerned feelings for people less fortunate than me. **	0.904	4.912 (4)	.296	0.307
I am often quite touched by things that I see happen.	0.870	8.082 (4)	.089	0.368
I would describe myself as a pretty soft-hearted person.	0.818	6.244 (4)	.182	0.303

Note: Factor loadings are the standardized loadings from the constrained model. Satorra-Bentler χ2 (df) is the change in χ2 value for the fully constrained model versus a model wherein the parameters for each item are freely estimated. P-value is the p-value associated with the difference between the models. dMACs are the degree of measurement non-equivalence for each item from the configural model with males serving as the reference group. For each study, the non-invariant item with the highest factor loading is marked by an asterisk. For the multidimensional model, anchor items that loaded onto the sympathy factor are bolded, and items that loaded onto the tenderness factor are in italics.

Table S33

AFIs and ΔAFIs for the measurement invariance analyses in Study I-2.

	df	*Robust CFI*	*Robust RMSEA*	*SRMR*
Unidimensional configural invariance model	22	0.999	0.038	0.019
Multidimensional configural invariance model*	20	0.999	0.035	0.017
ΔThreshold#	14	-0.0004	-0.0003	0.0
ΔPartial Threshold	13	-0.0002	-0.004	0.0
ΔMetric	5	-0.0002	0.001	0.001
ΔScalar	5	0.0	-0.0008	-0.0005
ΔStrict#	7	-0.001	0.012	0.009
ΔPartial strict	2	-0.001	0.014	0.009
Partial strict	45	0.998	0.046	0.026

Note: The configural model that was retained for analyses is marked by an asterisk. Models that failed invariance tests are marked by a hashtag. In the event that a partial invariant model is reported, the fit statistics represent the difference between the partial model, and the last non-invariant model.

Table S34

Ten largest standardized residual covariances for the multidimensional configural model in Study I-2.

Path	*Standardized residual covariance*	*Modification indice ($\Delta\chi^2$)*
Males		
"When I see someone being treated unfairly, I sometimes don't feel very much pity for them." ~~	-0.049	2.883

"I am often quite touched by things that I see happen."		
"When I see someone being treated unfairly, I sometimes don't feel very much pity for them." ~~ "When I see someone being taken advantage of, I feel kind of protective towards them."	0.047	0.001
“I often have tender, concerned feelings for people less fortunate than me.” ~~ “Other people's misfortunes do not usually disturb me a great deal.”	0.037	19.509
“I am often quite touched by things that I see happen.” ~~ “I would describe myself as a pretty soft-hearted person.”	0.036	0.091
"Other people's misfortunes do not usually disturb me a great deal." ~~ "When I see someone being taken advantage of, I feel kind of protective towards them."	-0.029	0.803
Females		
“I would describe myself as a pretty soft-hearted person.” ~~ “Other people's misfortunes do not usually disturb me a great deal.”	-0.039	0.056
“When I see someone being taken advantage of, I feel kind of protective towards them.” ~~ “Sometimes I don't feel very sorry for other people when they are having problems.”	-0.036	0.623
“I am often quite touched by things that I see happen.” ~~ “Sometimes I don't feel very sorry for other people when they are having problems.”	0.023	0.829
“Other people's misfortunes do not usually disturb me a great deal.” ~~ “When I see someone being taken advantage of, I feel kind of protective towards them.”	0.019	1.449
“I am often quite touched by things that I see happen.” ~~ “I often have tender, concerned feelings for people less fortunate than me.”	-0.014	0.185

Note: Modification indice is the $\Delta\chi^2$ after adding the covariance to the model. ~~ indicates a shared covariance between adjectives. We note that, when computing modification indices for

Study I-2, the results may not be trustworthy due to a correlation greater than 1 between the sympathy and tenderness factors.

Permutation test

We ran a permutation test by generating 1,000 randomly permuted datasets and setting maxSparse = 20. The results of the permutation test indicated that there was no evidence against the null of group discrepancy (*Scaled* $\chi^2(20) = 56.742$, $p = .332$), suggesting that the discrepancy was due to an overall approximation discrepancy in the baseline model, rather than a configural misspecification for men versus women.

Study J-1

Analysis code

Analyses for this study are included in an R script titled "Study_J_MI.R."

Model specification

Experimental manipulations and regressions

We regressed the experimental manipulations (0 = *Study 1*, 1 = *Study 2*) on the empathy factor for the unidimensional model, and, for the multidimensional model, on both the sympathy and tenderness factors.

Results

Table S35

Results of AOAA anchor selection analyses for Study J-1, and the dMACs for each configural model.

	Factor loading	*Satorra-Bentler χ2 (df)*	*p-value*	*dMAC*
Unidimensional model				
Sympathetic	0.823	7.676 (6)	.263	0.23
Compassionate	0.683	3.952 (6)	.683	0.22
Moved	0.859	4.092 (6)	.664	0.08
Softhearted*	0.970	9.430 (6)	.151	0.05
Tender	0.791	0.769 (6)	.993	0.09
Multidimensional model				
Sympathetic	0.830	6.530 (6)	.367	0.23
Compassionate	0.686	4.155 (6)	.656	0.22
Moved*	0.868	4.793 (6)	.571	0.08
*Softhearted**	0.986	4.886 (6)	.559	0.06
Tender	0.791	4.293 (6)	.637	0.10

Note: Factor loadings are the standardized loadings from the constrained model. Satorra-Bentler χ2 (df) is the change in χ2 value for the fully constrained model versus a model wherein the parameters for each item are freely estimated. P-value is the p-value associated with the difference between the models. dMACs are the degree of measurement non-equivalence for each item from the configural model with males serving as the reference group. For each study, the non-invariant item with the highest factor loading is marked by an asterisk. Items that showed evidence of non-invariance are marked by a hashtag. For the multidimensional model, anchor items that loaded onto the sympathy factor are bolded, and items that loaded onto the tenderness factor are in italics.

Table S36

AFIs and ΔAFIs for the measurement invariance analyses in Study J-1.

	df	*Robust CFI*	*Robust RMSEA*	*SRMR*
Unidimensional configural*	26	1.00	0.032	0.026
Multidimensional configural	20	1.00	0.033	0.025
ΔThreshold	20	0.0	-0.011	0.0
ΔMetric	4	-0.0001	0.002	0.0007
ΔScalar	4	0.0004	-0.013	0.0001
ΔStrict	5	0.0001	-0.011	0.002
Strict	59	1.00	0.0	0.029

Note: The configural model that was retained for analyses is marked by an asterisk.

Table S37

Ten largest residual correlations for the multidimensional configural model in Study J-1.

Path	*Residual correlation*	*Modification indice ($\Delta\chi^2$)*
Males		
Compassionate ~~ Sympathetic	0.07	6.340
Compassionate ~~ Softhearted	-0.06	4.858
Moved ~~ Sympathetic	-0.02	0.622
Tender ~~ Sympathetic	-0.01	0.015
Compassionate ~~ Moved	-0.01	0.00
Females		
Compassionate ~~ Sympathetic	0.08	1.389
Compassionate ~~ Softhearted	-0.06	1.219
Moved ~~ Tender	-0.02	0.003
Moved ~~ Sympathetic	-0.02	0.265
Tender ~~ Softhearted	0.01	0.129

Note: Modification indice is the $\Delta\chi^2$ after adding the covariance to the model. ~~ indicates a shared covariance between adjectives.

Study J-2

Analysis code

Analyses for this study are included in an R script titled “Study_J_MI.R.”

Model specification

Residual correlations

We included residual correlations between the three negatively worded items for both the unidimensional and multidimensional models: “Other people's misfortunes do not usually disturb me a great deal,” “Sometimes I don't feel very sorry for other people when they are having problems,” and “When I see someone being treated unfairly, I sometimes don't feel very much pity for them.”

Category collapsing

None of the males, and only two of the females (1.9% of all females), endorsed a response category less than “3” for the item “sympathetic”, so responses of “1”, “2”, or “3” for “sympathetic” were collapsed into a single response category of “3” for both men and women.

Results

Table S38

Results of AOAA anchor selection analyses for Study J-2, and the dMACs for each configural model.

	Factor loading	*Satorra-Bentler χ2 (df)*	*p-value*	*dMAC*
Unidimensional model				
Sometimes I don't feel very sorry for other people when they are having problems.	0.525	6.542 (3)	.088	0.37
When I see someone being taken advantage of, I feel kind of protective towards them.	0.621	4.035 (4)	.401	0.30
Other people's misfortunes do not usually disturb me a great deal.	0.525	5.591 (4)	.232	0.23
When I see someone being treated unfairly, I sometimes don't feel very much pity for them. #	0.523	10.016 (4)	.040*	0.25
I often have tender, concerned feelings for people less fortunate than me. *	0.741	4.715 (4)	.318	0.37
I am often quite touched by things that I see happen. #	0.729	12.705 (4)	.013*	0.50
I would describe myself as a pretty soft-hearted person. #	0.676	9.995 (4)	.041*	0.41

Sympathy-tenderness model				
Sometimes I don't feel very sorry for other people when they are having problems.	0.525	5.174 (3)	.160	0.36
When I see someone being taken advantage of, I feel kind of protective towards them. *	0.606	5.306 (4)	.257	0.25
Other people's misfortunes do not usually disturb me a great deal.	0.513	2.717 (4)	.606	0.20
When I see someone being treated unfairly, I sometimes don't feel very much pity for them.	0.503	2.493 (4)	.646	0.14
I often have tender, concerned feelings for people less fortunate than me. #	0.785	10.792 (4)	.029*	0.33
*I am often quite touched by things that I see happen. **	0.763	4.418 (4)	.352	0.47
I would describe myself as a pretty soft-hearted person.	0.706	2.591 (4)	.629	0.37

Note: Factor loadings are the standardized loadings from the constrained model. Satorra-Bentler χ2 (df) is the change in χ2 value for the fully constrained model versus a model wherein the parameters for each item are freely estimated. P-value is the p-value associated with the difference between the models. * = *p*-values < .05. dMACs are the degree of measurement non-equivalence for each item from the configural model with males serving as the reference group. For each study, the non-invariant item with the highest factor loading is marked by an asterisk. Items that showed evidence of non-invariance are marked by a hashtag. For the model, anchor items that loaded onto the sympathy factor are bolded, and items that loaded onto the tenderness factor are in italics.

Table S39

AFIs and ΔAFIs for the measurement invariance analyses in Study J-2.

	df	*Robust CFI*	*Robust RMSEA*	*Robust SRMR*
Unidimensional configural*	22	0.993	0.075	0.049
Multidimensional configural	20	0.993	0.076	0.047
ΔThreshold	13	0.0007	-0.018	0.0
ΔMetric#	6	-0.007	0.019	0.004

ΔPartial metric	4	0.0009	-0.007	0.001
ΔScalar	6	-0.002	0.004	0.002
ΔStrict	2	0.002	-0.011	0.002
Strict	52	0.994	0.042	0.055

Note: The configural model that was retained for analyses is marked by an asterisk. Models that failed invariance tests are marked by a hashtag. In the event that a partial invariant model is reported, the fit statistics represent the difference between the partial model, and the last non-invariant model.

Table S40

Ten largest residual covariances for the unidimensional configural model in Study J-2.

Path	*Residual covariance*	*Modification indice ($\Delta\chi^2$)*
Males		
"I would describe myself as a pretty soft-hearted person." ~~ "Other people's misfortunes do not usually disturb me a great deal."	0.18	0.637
"I would describe myself as a pretty soft-hearted person." ~~ "When I see someone being treated unfairly, I sometimes don't feel very much pity for them."	-0.11	0.236
"I am often quite touched by things that I see happen." ~~ "Other people's misfortunes do not usually disturb me a great deal."	-0.10	0.037
"I am often quite touched by things that I see happen." ~~ "Sometimes I don't feel very sorry for other people when they are having problems."	-0.07	0.510
"I am often quite touched by things that I see happen." ~~ "I often have tender, concerned feelings for people less fortunate than me."	0.07	2.759
Females		
"I would describe myself as a pretty soft-hearted person." ~~ "When I see someone being taken advantage of, I feel kind of protective towards them. "	-0.13	1.724
"When I see someone being taken advantage of, I feel kind of protective towards them." ~~ "I often have tender, concerned feelings for people less fortunate than me."	0.11	11.763
"I am often quite touched by things that I see happen." ~~ "When I see someone being treated unfairly, I sometimes don't feel very much pity for them."	-0.11	1.581
"When I see someone being taken advantage of, I feel kind of protective towards them." ~~	-0.05	0.247

"Sometimes I don't feel very sorry for other people when they are having problems."		
"I would describe myself as a pretty soft-hearted person." ~~ "I am often quite touched by things that I see happen."	0.05	0.010

Note: Modification indice is the $\Delta\chi^2$ after adding the covariance to the model. ~~ indicates a shared covariance between adjectives.

Permutation test

We attempted to run a permutation test by generating 1,000 randomly permuted datasets and setting maxSparse = 20. However, due to excessive sparseness in response categories for several of the items, we could not complete the analysis.

Study K-1

Analysis code

Analyses for this study are included in an R script titled "Study_K_MI.R."

Model specification

Experimental manipulations and regressors

This study did not include any experimental manipulations that needed to be controlled for in analyses.

Results

Table S41

Results of AOAA anchor selection analyses for Study K-1, and the dMACs for each configural model.

	Factor loading	*Satorra-Bentler χ2 (df)*	*p-value*	*dMAC*
Unidimensional model				
Sympathetic*	0.937	4.999 (6)	.544	0.306
Compassionate	0.912	2.077 (6)	.913	0.343
Empathic	0.879	11.636 (6)	.071	0.416
Moved	0.877	8.701 (6)	.191	0.260
Concerned	0.718	12.338 (6)	.055	0.068
Warm	0.772	4.295 (6)	.637	0.174
Softhearted	0.931	7.029 (6)	.318	0.310
Tender	0.848	7.90 (6)	.246	0.296
Multidimensional model				
Sympathetic*	0.942	8.957 (6)	.176	0.308
Compassionate	0.915	1.585 (6)	.954	0.344
Empathic	0.883	11.759 (6)	.068	0.416
Moved	0.885	6.243 (6)	.397	0.261

Concerned	0.724	9.026 (6)	.172	0.069
Warm	0.778	3.017 (6)	.807	0.175
*Softhearted**	0.947	10.841 (6)	.093	0.312
Tender	0.855	10.576 (6)	.102	0.297

Note: Factor loadings are the standardized loadings from the constrained model. Satorra-Bentler χ2 (df) is the change in χ2 value for the fully constrained model versus a model wherein the parameters for each item are freely estimated. P-value is the p-value associated with the difference between the models. dMACs are the degree of measurement non-equivalence for each item from the configural model with males serving as the reference group. For each study, the non-invariant item with the highest factor loading is marked by an asterisk. Items that showed evidence of non-invariance are marked by a hashtag. For the multidimensional model, anchor items that loaded onto the sympathy factor are bolded, and items that loaded onto the tenderness factor are in italics.

Table S42

AFIs and ΔAFIs for the measurement invariance analyses in Study K-1.

	df	*Robust CFI*	*Robust RMSEA*	*SRMR*
Unidimensional configural	40	0.998	0.120	0.043
Multidimensional configural	38	0.998	0.113	0.043
ΔThreshold	32	0.0	-0.029	0.0
ΔMetric	6	0.0	-0.003	0.0
ΔScalar	6	0.0	-0.001	0.0
ΔStrict	8	-0.0004	0.003	0.005
ΔPartial strict#	6	-0.0004	0.004	0.005
Partial strict	88	0.997	0.084	0.047

Note: The configural model that was retained for analyses is marked by an asterisk. Non-invariant models are marked by a hashtag.

Table S43

Ten largest standardized residual covariances for the multidimensional configural model in Study K-1.

Path	*Standardized residual covariance*	*Modification indice ($\Delta\chi^2$)*
Males		
Warm ~~ Concerned	-0.10	4.414
Moved ~~ Compassionate	-0.08	7.982
Moved ~~ Empathic	-0.07	5.154
Warm ~~ Moved	0.07	5.924
Warm ~~ Tender	0.07	4.591
Females		
Warm ~~ Concerned	-0.16	7.796
Moved ~~ Empathic	-0.11	10.314
Concerned ~~ Compassionate	0.08	6.381
Warm ~~ Empathic	-0.07	3.737
Sympathetic ~~ Moved	-0.07	6.007

Note: Modification indice is the $\Delta\chi^2$ after adding the covariance to the model. ~~ indicates a shared covariance between adjectives.

Permutation test

We ran a permutation test by generating 1,000 randomly permuted datasets and setting maxSparse = 20. The results of the permutation test indicated that there was no evidence against the null of group discrepancy (*Scaled* χ^2(38) = 361.479, *p* = .142), suggesting that the discrepancy was due to an overall approximation discrepancy in the baseline model, rather than a configural misspecification for men versus women.

Study K-2

Model specification

Residual correlations

We included residual correlations between the three negatively worded items for both the one- and multidimensional models: "Other people's misfortunes do not usually disturb me a great deal," "Sometimes I don't feel very sorry for other people when they are having problems," and "When I see someone being treated unfairly, I sometimes don't feel very much pity for them."

Experimental manipulations and regressions

This study did not include any experimental manipulations that needed to be controlled for in analyses.

Category collapsing

None of the females, and only four of the males (2.3% of all males), endorsed a response category less than "2" for the item "I often have tender, concerned feelings for people less fortunate than me," so responses of "1" or "2" for this items were collapsed into a single response category of "2" for both men and women.

Results

Table S44

Results of AOAA anchor selection analyses for Study K-2, and the dMACs for each configural model.

	Factor loading	*Satorra-Bentler χ2 (df)*	*p-value*	*dMAC*
Unidimensional model				
Sometimes I don't feel very sorry for other people when they are having problems.	0.513	6.641 (4)	.156	0.375
When I see someone being taken advantage of, I feel kind of protective towards them.	0.796	5.247 (4)	.263	0.213

Other people's misfortunes do not usually disturb me a great deal.	0.655	1.788 (4)	.775	0.373
When I see someone being treated unfairly, I sometimes don't feel very much pity for them.	0.651	2.843 (4)	.585	0.080
I often have tender, concerned feelings for people less fortunate than me.*	0.887	1.329 (3)	.722	0.401
I am often quite touched by things that I see happen.	0.781	1.090 (4)	.896	0.446
I would describe myself as a pretty soft-hearted person.	0.858	3.960 (4)	.411	0.278
Sympathy-tenderness model				
Sometimes I don't feel very sorry for other people when they are having problems.	0.553	6.953 (4)	.138	0.358
When I see someone being taken advantage of, I feel kind of protective towards them.*	0.841	7.149 (4)	.128	0.203
Other people's misfortunes do not usually disturb me a great deal.	0.707	1.876 (4)	.759	0.355
When I see someone being treated unfairly, I sometimes don't feel very much pity for them.	0.700	2.519 (4)	.641	0.105
*I often have tender, concerned feelings for people less fortunate than me.**	0.900	0.653 (3)	.884	0.400
I am often quite touched by things that I see happen.	0.793	2.092 (4)	.719	0.443

I would describe myself as a pretty soft-hearted person.	0.867	4.173 (4)	.383	0.276

Note: Factor loadings are the standardized loadings from the constrained model. Satorra-Bentler χ2 (df) is the change in χ2 value for the fully constrained model versus a model wherein the parameters for each item are freely estimated. P-value is the p-value associated with the difference between the models. dMACs are the degree of measurement non-equivalence for each item from the configural model with males serving as the reference group. For each study, the non-invariant item with the highest factor loading is marked by an asterisk. For the model, anchor items that loaded onto the sympathy factor are bolded, and items that loaded onto the tenderness factor are in italics.

Table S45

AFIs and ΔAFIs for the measurement invariance analyses in Study K-2.

	df	*Robust CFI*	*Robust RMSEA*	*Robust SRMR*
Unidimensional configural	22	0.997	0.076	0.034
Multidimensional configural*	20	0.998	0.065	0.030
ΔThreshold	13	-0.0001	-0.013	0.0
ΔMetric	5	0.0006	-0.01	0.0
ΔScalar	5	-0.0004	0.002	0.0002
ΔStrict#	7	-0.003	0.02	0.011
ΔPartial strict#	4	-0.003	0.02	0.011
Scalar	43	0.998	0.044	0.030

Note: The configural model that was retained for analyses is marked by an asterisk. Models that failed invariance tests are marked by a hashtag. In the event that a partial invariant model is reported, the fit statistics represent the difference between the partial model, and the last non-invariant model.

Table S46

Ten largest standardized residual covariances for the multidimensional configural model in Study K-2.

Path	*Standardized residual covariance*	*Modification indice ($\Delta\chi^2$)*
Males		
"I am often quite touched by things that I see happen." ~~ "When I see someone being treated unfairly, I sometimes don't feel very much pity for them."	-0.09	3.572
"I am often quite touched by things that I see happen." ~~ "I would describe myself as a pretty soft-hearted person."	0.07	10.964
"When I see someone being treated unfairly, I sometimes don't feel very much pity for them." ~~	-0.07	1.903

"I would describe myself as a pretty soft-hearted person."		
"Other people's misfortunes do not usually disturb me a great deal." ~~ "When I see someone being taken advantage of, I feel kind of protective towards them."	-0.07	3.537
"I am often quite touched by things that I see happen." ~~ "I often have tender, concerned feelings for people less fortunate than me."	-0.06	5.810
Females		
"When I see someone being taken advantage of, I feel kind of protective towards them." ~~ "I would describe myself as a pretty soft-hearted person."	-0.07	0.092
"I would describe myself as a pretty soft-hearted person." ~~ "Sometimes I don't feel very sorry for other people when they are having problems."	0.05	5.225
"Sometimes I don't feel very sorry for other people when they are having problems." ~~ "When I see someone being taken advantage of, I feel kind of protective towards them."	-0.04	0.750
"I am often quite touched by things that I see happen." ~~ "When I see someone being treated unfairly, I sometimes don't feel very much pity for them."	-0.03	0.019
"When I see someone being taken advantage of, I feel kind of protective towards them." ~~ "I am often quite touched by things that I see happen."	0.03	10.369

Note: Modification indice is the $\Delta\chi^2$ after adding the covariance to the model. ~~ indicates a shared covariance between adjectives.

Permutation test

We attempted to conduct a permutation test by generating 1,000 randomly permuted datasets and setting maxSparse = 20. However, we could not complete analyses due to excessive empty categories during permutation, so we updated maxSparse = 60 and re-ran the permutation test. The results of the permutation test indicated that there was no evidence against the null of group discrepancy (*Scaled* $\chi^2(20) = 63.979$, $p = .280$), suggesting that the discrepancy was due to an overall approximation discrepancy in the baseline model, rather than a configural misspecification for men versus women.

Summary of residual covariances for the failed configural trait models

We examined the modification indices to better understand the misspecifications that may have caused the configural models to fail. For Study C, the largest residual covariance for both males and females was the residual covariance between the items “sympathetic” and “compassionate” (*cov*s < -7.08, $\Delta\chi^2$s > 3.44). For Study D, the largest residual correlation for males was “empathic” with “concerned” ($r = -0.03$, $\Delta\chi^2 = 1.05$). For females, the largest residual correlation was “compassionate” with “tender” ($r = 0.06$, $\Delta\chi^2 = 9.20$). For Study F, the largest residual correlation was between “tender” and “sympathetic” for males ($r = -0.02$, $\Delta\chi^2 = 0.09$), and “softhearted” and “empathic” for females ($r = -0.03$. $\Delta\chi^2 = 0.63$)[2]. For Study I-1, the largest standardized residual covariance for both males and females was between the items “sympathetic” and “moved” (*cov*s < -0.034, $\Delta\chi^2$s > 6.691). For Study K-1, the largest standardized residual covariance for both males and females was between the items “warm” and “concerned” (*cov*s < -0.10, $\Delta\chi^2$s > 4.414). Overall, there was no consistent pattern across studies, suggesting that sources of misfit may be sample-specific. We did not add in any residual covariances to improve the fit of the models.

Summary of residual covariances for the failed configural trait models

We examined the modification indices for the failed configural model to better understand the cause of the misspecifications. For Study D-2, the largest residual covariance for males was shared by the items “I am often quite touched by things that I see happen” and “I often have tender, concerned feelings for people less fortunate than me” ($cov = 7.84$, $\Delta\chi^2 = 1.50$); for females, the largest residual covariance was “When I see someone being taken advantage of, I feel kind of protective towards them” with “When I see someone being treated unfairly, I sometimes don't feel very much pity for them” ($cov = 6.84$, $\Delta\chi^2 = 10.77$).

For Study G, the largest residual correlation for males was shared by the items, “I am often quite touched by things that I see happen” and “Other people's misfortunes do not usually disturb me a great deal” ($r = -0.09$, $\Delta\chi^2 = 1.50$); for females, the largest residual correlation was “I am often quite touched by things that I see happen” with “When I see someone being taken advantage of, I feel kind of protective towards them” ($r = -0.13$, $\Delta\chi^2 = 3.70$).

For Study I-2, the largest residual covariance for males was shared by the items, “When I see someone being treated unfairly, I sometimes don't feel very much pity for them” with “I am often quite touched by things that I see happen” ($cov = -0.049$, $\Delta\chi^2 = 2.883$). For females, the largest covariance was shared between “I would describe myself as a pretty soft-hearted person” and “Other people's misfortunes do not usually disturb me a great deal” ($cov = -0.039$, $\Delta\chi^2 = 0.056$).

For Study J-2, the largest residual covariance for males was shared by the items, “I would describe myself as a pretty soft-hearted person” with “Other people's misfortunes do not usually disturb me a great deal” ($cov = 0.18$, $\Delta\chi^2 = 0.637$). For females, the largest covariance was shared

[2] For some of the failed configural models we report residual correlations, while for others we report residual covariances. Different metrics were used for each study, because some models included structural specifications that change the metric used to report residuals. See Supplemental Materials for details about the model specification in each study.

between “I would describe myself as a pretty soft-hearted person.” with “When I see someone being taken advantage of, I feel kind of protective towards them” (*cov* = -0.13, $\Delta\chi^2$ = 1.724).

Finally, for Study K-2, the largest residual covariance for males was shared by the items, “I am often quite touched by things that I see happen” with “When I see someone being treated unfairly, I sometimes don't feel very much pity for them.” (*cov* = -0.09, $\Delta\chi^2$ = 3.572). For females, the largest covariance was shared between “When I see someone being taken advantage of, I feel kind of protective towards them” with “I would describe myself as a pretty soft-hearted person” (*cov* = -0.07, $\Delta\chi^2$ = 0.092).

Summary of measurement invariance violations for state empathic concern

In Study I-1 we encountered two violations of measurement invariance. First, we found the threshold invariant model was non-invariant (*Scaled* $\chi^2(20)$ = 43.347, *p* = .002). We inspected the modification indices associated with the scalar invariance model. Modification indices revealed that two of the thresholds had *p*-values < .05: the fourth, and sixth, thresholds for “softhearted,” and “concerned.” We attempted to fit a partial threshold invariant model that freely estimated the thresholds for the two items in each group and compared the fit of the partial thresholds and configural invariant models. The partial threshold invariant model did not fit significantly worse than the metric invariant model (*Satorra-Bentler* $\chi^2(18)$ = 26.804 *p* = .083), indicating that partial threshold invariance could be achieved.

Second, we found that the strict invariant model for study I-1 was non-invariant (*Scaled* $\chi^2(5)$ = 17.846, *p* = .003). Modification indices revealed that the residual for the items “moved” and “tender” had *Wald* χ^2 > 10. We fit a partial strict invariance model that freely estimating the residuals for “moved” and “tender” in each group would significantly improve model fit, and we compared the fit of the partial scalar and partial strict invariant models. The partial strict invariant model still fit significantly worse than the partial scalar invariant model (*Satorra-Bentler* $\chi^2(3)$ = 10.708, *p* = .013). We then terminated invariance testing.

In Study K-1, we found that the strict invariant model was non-invariant (*Scaled* $\chi^2(8)$ = 18.827, *p* = .016). Modification indices revealed that the residual for the items “moved” and “tender” had Wald χ^2 > 10. We fit a partial strict invariance model that freely estimating the residuals for “moved” and “tender” in each group would significantly improve model fit, and we compared the fit of the partial scalar and partial strict invariant models. The partial strict invariant model still fit significantly worse than the scalar invariant model (Satorra-Bentler $\chi^2(6)$ = 14.12, *p* = .028). We then terminated invariance testing.

Summary of measurement invariance violations for trait empathic concern

For Study D-2, the strict invariant model was non-invariant (*Satorra-Bentler* $\chi^2(6)$ = 28.408, *p* < .001). Modification indices revealed that the residual variances associated with three items had *Wald* χ^2 > 10: “Other people's misfortunes do not usually disturb me a great deal,” “Sometimes I don't feel very sorry for other people when they are having problems,” and “When I see someone being treated unfairly, I sometimes don't feel very much pity for them.” We attempted to fit a partial strict invariance model that freely estimated the residual variances for the three items in each group, and compared the fits of the partial strict and scalar invariant models, but found that the partial strict invariant model still had significantly worse fit than the scalar invariant model (*Satorra-Bentler* $\chi^2(4)$ = 17.857, *p* = .001). We terminated invariance testing and estimated the bias in the effect size caused by inspecting the dMACs for items from the scalar invariant model. All dMACs were vanishingly small (*dMACs* < .001), indicating that differences in the structure of residuals negligibly biased observed differences in empathic concern.

For study G, we found that the scalar invariant model was non-invariant (*Satorra-Bentler* $\chi^2(5) = 11.40$, $p = .044$). Modification indices indicated that freeing the intercept for the item “Other people's misfortunes do not usually disturb me a great deal” would significantly improve model fit ($\chi^2(1) = 14.84$, $p = .001$). We fit a partial scalar invariance model that freely estimated the intercept for “Other people's misfortunes do not usually disturb me a great deal,” and compared it to the metric invariant model. The fit of the partial scalar invariant model was not significantly different than the fit of the metric invariant model (*Satorra-Bentler* $\chi^2(4) = 3.493$, $p = .479$), and all items from the partial scalar invariant model had dMACs approaching zero (*dMACs* < .001).

For Study I-2, we found multiple violations of measurement invariance. First, we found threshold non-invariance (*Satorra-Bentler* $\chi^2(14) = 26.922$, $p = .020$). Freeing the fourth thresholds for the item "I would describe myself as a pretty soft-hearted person” significantly improved model fit (*Satorra-Bentler* $\chi^2(1) = 3.989$, $p = .046$). The fit of the partial threshold invariant model was not statistically different from the fit of the configural model (*Satorra-Bentler* $\chi^2(13) = 19.304$, $p = .114$).

We additionally found that, upon comparing the scalar invariant model to the strict invariant model in Study I-2, the strict invariant model was non-invariant (*Satorra-Bentler* $\chi^2(7) = 25.126$, $p < .001$). Modification indices indicated that freeing the residuals for five of the items would significantly improve model fit (*Wald* $\chi^2 s$ > 10.285): "When I see someone being taken advantage of, I feel kind of protective towards them”; "Other people's misfortunes do not usually disturb me a great deal"; "When I see someone being treated unfairly, I sometimes don't feel very much pity for them”; "I am often quite touched by things that I see happen”; and “I would describe myself as a pretty soft-hearted person.” We freed the residuals for these five items, and tested the fit of the partial strict invariant model against the scalar invariant model, but the partial strict invariant model remained non-invariant (*Satorra-Bentler* $\chi^2(2) = 7.212$, $p = .027$); all items from the partial strict invariant model had dMACs approaching zero (*dMACs* < .001).

For Study J-2, we found the metric invariant model was non-invariant (*Satorra-Bentler* $\chi^2(6) = 19.08$, $p = .004$). Modification indices indicated that freeing the loadings for the item “I am often quite touched by things that I see happen” would significantly improve model fit ($\chi^2(1) = 14.84$, $p = .001$). However, the fit of the partial metric invariant model remained significantly different than the fit of the threshold invariant model (*Satorra-Bentler* $\chi^2(5) = 11.206$, $p = .047$). We then freed the loading for the item with the next largest modification index, “I would describe myself as a pretty soft-hearted person,” and compared it to the threshold invariant model. The fit of the partial scalar invariant model was not significantly different than the fit of the metric invariant model (*Satorra-Bentler* $\chi^2(4) = 2.249$, $p = .690$). We then inspected the dMACS from the partial metric invariant model, and found medium-sized dMACs for “I am often quite touched by things that I see happen” (*dMAC* = 0.37) and “I would describe myself as a pretty soft-hearted person” (*dMAC* = 0.33).

Finally, for Study K-2, we found the strict invariant model was non-invariant (*Satorra-Bentler* $\chi^2(7) = 28.885$, $p < .001$). Modification indices indicated that freeing the residuals for the items “Sometimes I don't feel very sorry for other people when they are having problems”, “When I see someone being treated unfairly, I sometimes don't feel very much pity for them", and "I would describe myself as a pretty soft-hearted person" would significantly improve model fit (*Wald* χ^2*s* > 12.043). We freed the residuals for the three items in each group, refit the strict invariant model, and compared the partial strict invariant model against the scalar invariant model. The fit of the partial strict invariant model remained significantly different than the fit of the scalar invariant model (*Satorra-Bentler* $\chi^2(4) = 16.506$, $p = .002$). We then inspected the dMACS from the partial strict invariant model, and found the dMACs were vanishingly small (*dMACs* < 0.001).

Summary of divergences between the $\Delta\chi^2$ and ΔCFI

For each study, we focus on reporting the χ^2 to guide our decisions about whether or not invariance had been achieved. Here, we report any instances in which our decisions may have diverged for using the $\Delta\chi^2$ criterion, compared to ΔCFI. Our decision criteria was to fail to reject any model for which the ΔCFI ≤ -0.001 (Cheung & Rensvold, 2002).

For Study E, the ΔCFI criterion indicates that the scalar and strict invariance models should be rejected. In contrast, the $\Delta\chi^2$ criterion indicates that the models should not be rejected. For all other measurement invariance models in which the $\Delta\chi^2$ criterion indicates that the models should be rejected, the ΔCFI criterion indicates that the models should *not* be rejected (please see Tables 4 and 8 for the instances in which the $\Delta\chi^2$ criterion indicates that the models should be rejected). In other words, every model that was rejected according to $\Delta\chi^2$ was accepted according to ΔCFI.