



Reliability of the empathy selection task, a novel behavioral measure of empathy avoidance

Amanda M. Ferguson^{1,2} · Michael Inzlicht^{3,4}

Accepted: 26 June 2022 / Published online: 22 August 2022
© The Psychonomic Society, Inc. 2022

Abstract

The empathy selection task is a novel behavioral paradigm designed to assess an individual's willingness to engage in empathy. Work with this task has demonstrated that people prefer to avoid empathy when some other activity is available, though individual differences that might predict performance on this task have been largely unexamined. Here, we assess the suitability of the empathy selection task for use in individual difference and experimental research by examining its reliability within and across testing sessions. We compare the reliability of summary scores on the empathy selection task (i.e., proportion of empathy choices) as an individual difference metric to that of two commonly used experimental tasks, the Stroop error rate and go/no-go commission rate. Next, we assess systematic changes at the item/trial level using generalized multilevel modeling which considers participants' individual performance variation. Across two samples ($N = 89$), we find that the empathy selection task is stable between testing sessions and has good/substantial test-retest reliability ($ICCs = .65$ and $.67$), suggesting that it is comparable or superior to other commonly used experimental tasks with respect to its ability to consistently rank individuals.

Keywords Reliability · Individual differences · Empathy · Behavioral task

Most of us experience empathy many times per day (Depow et al., 2021). Yet, our willingness to empathize depends critically on the situation. A parent is more likely to empathize with their child than with another child at the playground, not for a lack of empathic ability but because their role as a parent makes empathizing with their child more motivating for them (e.g., see Ferguson et al., 2020). In other words, one's propensity and capacity for empathy can vary (Keyesers & Gazzola, 2014); the extent to which someone is capable of empathy is not necessarily related to their willingness to engage it in any particular situation. Empathy propensity is most often measured via self-report, but a novel behavioral

paradigm, called the empathy selection task (Cameron et al., 2019), has recently been created. Evidence related to the validity of the empathy selection task has been presented elsewhere (Cameron et al., 2019), but its reliability has yet to be comprehensively examined. Given that empathy is often studied in relation to psychological factors that vary across individuals (e.g., personality, interpersonal functioning, psychological health), the present work examines the reliability of the empathy selection task for use in within- and between-person research designs.

Empathy has been studied for decades, and for good reason. Dispositional empathy is associated with interpersonal closeness and relationship satisfaction (Cramer & Jowett, 2010), it predicts prosocial action (Depow et al., 2021; Pavey et al., 2012; Van Lange, 2008), and it facilitates the acceptance of social support in times of stress (Cosley et al., 2010). On the other hand, there are costs to feeling with those around us. Empathy feels effortful (Cameron et al., 2019; Ferguson et al., 2021), and it can lead to unfair (Batson et al., 1995) and even unethical (Pierce et al., 2013) behavior. Studying these processes requires that we define and measure empathy, and attempts to do so have demonstrated that it is a multifaceted process (e.g., see Batson, 2009; Bernhardt & Singer, 2012;

✉ Amanda M. Ferguson
am.ferguson@mail.utoronto.ca

¹ Graduate Department of Psychological Clinical Science, University of Toronto, Toronto, ON, Canada

² MRC Cognition and Brain Sciences Unit, University of Cambridge, Cambridge, United Kingdom

³ Department of Psychology, University of Toronto, Toronto, ON, Canada

⁴ Rotman School of Management, University of Toronto, Toronto, ON, Canada

Davis et al., 1994; Decety & Jackson, 2004; Preston & De Waal, 2002), with distinct but overlapping components: an affective component (e.g., experience sharing; feeling what someone else feels), a cognitive component (e.g., perspective taking; putting yourself in someone else's shoes), and a motivational component (e.g., compassion; having the intention to respond kindly to another person's distress; Decety & Jackson, 2004).

The importance of reliable measures

In the context of individual difference research, an instrument's reliability refers to its ability to consistently *rank* individuals at two or more time points. That is, a measure that consistently ranks person A above person B on some trait can be said to be reliable, even if both person A and B are in the top 1% of that trait. This is commonly assessed by calculating the intraclass correlation coefficient (ICC), which represents the proportion of the total variance in the data attributed to variance between individuals across time points.

In experimental psychology research, reliability is often understood as the extent to which a task can consistently *produce the same effect* across participants (e.g., the Stroop task reliably elicits a reaction time cost on incongruent trials; Hedge et al., 2018). Minimal between-participant variability on these tasks makes them especially useful for researchers, since nearly all participants will demonstrate the effect of interest (i.e., nearly everyone shows a Stroop effect; Haaf & Rouder, 2019; MacLeod, 1991), and the within-person variation of the effect can be examined with precision under different environmental conditions. However, tasks which produce very little between-participant variability will also be unlikely to consistently rank individuals, since the margin of difference between participants is minimized by design (for a fuller discussion of this idea, see Hedge et al., 2018).

Hedge et al. (2018) report a range of test-retest reliabilities across commonly used experimental task indices, the majority of which demonstrated values of $ICC = .7$ or below. Some task indices (e.g., spatial-numerical association of response code/SNARC effect RT) demonstrated values below $ICC = .4$. An ICC value of $.4$ means that, across time points, the variance accounted for by individuals amounts to only 40% of the total variance in the data, while the remaining 60% is accounted for by error variance and variance across testing sessions¹. A researcher interested in examining scores on this task as a factor that distinguishes between individuals within a population (i.e., between-subject variance) will therefore be at a disadvantage, since the variation across individuals explains

less than half the total variance available to them. In this way, certain experimental tasks might be particularly ill-suited for individual difference research (Hedge et al., 2018) as any correlation between these measures and another measure could be a function of the task's inability to consistently rank individuals (i.e., low reliability), a true lack of relation (or low relation) between constructs, or combination of the two (Dang, King, & Inzlicht, 2020).

This is an important area of concern for those interested in individual differences associated with performance on various experimental tasks. This issue is highlighted by several large-scale individual difference studies examining correlations among tasks of inhibitory control (e.g., Stroop task and flanker task), which have demonstrated that most do not exceed $.2$ in value and some have no demonstrated correlation at all (Hall, 2012; Hedge et al., 2018; Rey-Mermet et al., 2018). Nonetheless, since it is often theoretically meaningful to investigate individual differences in behavioral tasks, we must first assess a task's ability to consistently rank individuals.

Measuring empathy

Researchers interested in trait empathy (and individual differences in trait empathy) have several well-studied and psychometrically sound self-report questionnaires to choose from (e.g., see Neumann et al., 2015 for review). Behavioral measures of empathy are also available, many of which measure empathy ability and/or accuracy. For example, in the Reading the Mind in the Eyes test (Baron-Cohen, Jolliffe, Mortimore, & Robertson, 1997), participants are shown photographs of a person's eyes, and are asked to which of the presented words best describes what the person is thinking or feeling. Interestingly, a recent meta-analysis suggested only a weak relation ($r = .10$) between self-reported and behaviorally measured empathy (Murphy & Lilienfeld, 2019), which may be indicative of a true (if small) relation between self-reported and behavioral measures of empathy, an inability of the measures to consistently rank individuals (i.e., low reliability), or both.

The empathy selection task The empathy selection task (Cameron et al., 2019) was developed alongside a growing interest in evidence that changing people's motivations to empathize can shape empathic outcomes (Keysers & Gazzola, 2014; Zaki, 2014). A subjective value-based decision-making framework for understanding motivated empathy suggests people are weighing the costs of empathizing against offsetting rewards in any given situation. While previous work has established that material costs (e.g., money or time; Andreoni, Rao, & Trachtman, 2017; Shaw, Batson, & Todd, 1994) and emotional costs (e.g., vicarious distress; Cameron, Harris, & Payne, 2016; Cameron & Payne, 2011) can

¹ In ICC calculations for absolute agreement. Session variance is omitted from the ICC calculations for consistency agreement.

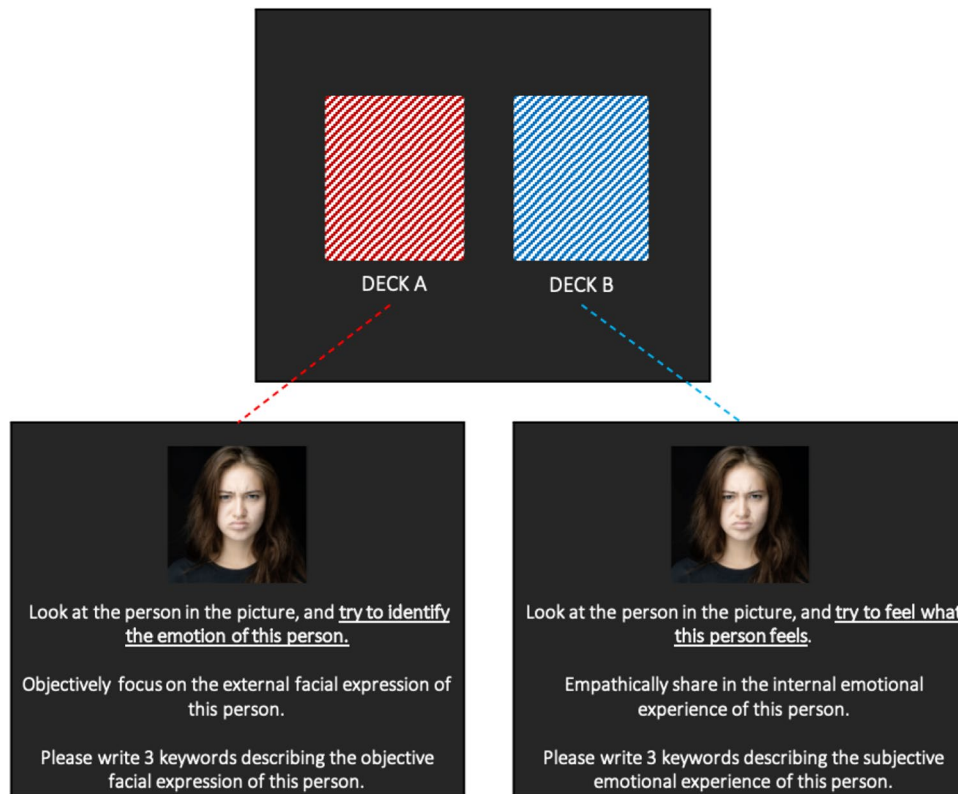


Fig. 1 Visualization of trial procedure for the empathy selection task. On each trial, participants see two decks of cards and are asked to choose between them. After choosing, they are shown an image and

given one of two sets of instructions. One deck will lead to instructions for the empathy task (here, DECK B), and the other to an alternative task (here, DECK A)

deter empathy behavior, the cognitive costs of empathizing have only recently been examined. Empathy entails a level of uncertainty and error—by definition, we have less information about the internal experiences of other people than we do about our own and making inferences about other's experience requires directed attention and working memory (Lin, Keysar, & Epley, 2010). Similarly, we run the risk of getting empathy wrong (e.g., misinterpreting cues; expecting one emotion and witnessing another), and past work has shown that concerns about making errors impacts perceptions of effort (Dunn, Inzlicht, & Risko, 2019).

The empathy selection task was developed to examine one specific cost that can affect our behavior: the *cognitive* work involved in empathy. Based on established effort avoidance tasks (i.e., the demand selection task; Kool et al., 2010), the empathy selection task uses the logic of behavioral economics to assesses one's willingness to engage in empathy. This form of empathy regulation (i.e., via situation selection, whereby people choose situations to enter based upon the emotions they want to feel; Gross, 2013) is akin to experiences in everyday life, as people often choose to avoid (or not) environmental cues based on their predictions of the empathy demands that may arise (e.g., avoiding a donation

solicitor on the street). The task was designed to capture the regulation of experience/emotion sharing, and not the other facets of perspective taking or compassion. However, recent work on people's experience of empathy in everyday life demonstrates that emotion sharing, compassion, and perspective taking occur together around 75% of the time and rarely occur in isolation (Depow et al., 2021). Given these findings, it is reasonable to assume that individuals are not only engaging in experience/emotion sharing on the empathy selection task, but also (and perhaps to a lesser extent, given the explicit instructions related to experience/emotion sharing) compassion and perspective taking.

The empathy selection task is a behavioral paradigm, wherein participants make a series of binary choices between completing an empathy activity (e.g., empathizing with a stranger) or a comparable activity that does not require empathy (e.g., labeling the emotional expression of a stranger; see Fig. 1)². The nature and complexity of the

² Cameron et al. (2019) modified different task features to address different alternative explanations of empathy avoidance, but choosing between empathy and an alternative action is the core feature of the empathy selection task.

alternative, non-empathy activity offered on the empathy selection task is crucial component of the task design, and much work has been done to establish appropriate options (see Cameron et al., 2019). The task is designed so that individuals will, after gaining some experience with each activity, determine which is easiest and choose it more often. An alternative/non-empathy activity that is comparably fun, easy, or that takes less time to complete, will be attractive to participants regardless of the costs of empathizing (Kool et al., 2010). The most precise versions of the task solicit identical emotion information across activities, and only vary whether participants are also instructed to engage in empathy (e.g., Studies 3 and 6 from Cameron et al., 2019).

Across several different samples and paradigms (e.g., see Cameron et al., 2019; Cameron & Inzlicht, 2020; Ferguson et al., 2020; Scheffer et al., 2022), results have suggested that people prefer to avoid empathizing with strangers, and that these preferences are driven by perceptions of cognitive costs (e.g., mental demand, negative affect, and inefficaciousness) associated with empathizing. Empathy avoidance on the empathy selection task has been demonstrated regardless of whether people are asked to empathize with pleasant or unpleasant emotions, and when the alternative task also includes emotional content (Cameron et al., 2019). Perceptions of cognitive costs are consistently associated with empathy avoidance, and people are less likely to choose the empathy deck as time on task increases. Similarly, manipulating people's sense of self-efficacy (via false feedback) during the empathy trials increases people's willingness to opt into empathy on the task. Ferguson et al. (2020) demonstrated that people were more willing to engage in empathy with a self-nominated loved-other than they were a stranger on the task, and that they reported having more self-efficacy when empathizing with loved-others versus strangers.

Many of the rewards (e.g., affiliation) and costs (e.g., helping) that often accompany empathy in the real world are intentionally removed from the empathy selection task (see Ferguson et al., 2021 for on this topic). This allows for more precise evaluation of the construct being assessed (i.e., effortful empathy approach/avoidance behavior), but also limits the suitability of the task for use in different contexts. That is, the empathy selection task is a finely sharpened tool which can be used to support carefully specified experimental hypotheses related to effortful emotion sharing (i.e., via situation selection), but it is not appropriate for use as a general measure of "empathy," since the complexity of the process of empathy is stripped away by design.

In the context of individual difference research, trial-level choices on the empathy selection task can be averaged into a summary score of task performance (i.e., proportion of empathy choices). As mentioned, this score does not index general or broadly defined empathy propensity, preference, or ability. Instead, summary scores on the empathy selection

task can be understood to index one's *tendency to avoid cognitive costs in the context of emotion sharing with strangers*. Importantly, some or even most of the variance on the task might reflect general effort avoidance which is nonspecific to empathy, since deck preferences can develop entirely based on effort-related properties of the non-empathy deck (and not at all because of empathy-related properties of the empathy deck). For example, individuals who habitually engage in little effort while completing tasks and surveys often produce different mean scores than do higher-effort respondents (Huang et al., 2015). Since the empathy selection task is inherently a measure of effort sensitivity (in the context of empathy), associations between summary scores and other variables may often be driven partly or heavily by whether someone is a high- or low-effort responder, rather than by some other theoretically plausible trait (e.g., see Huang et al., 2015).

In short, behavior on the empathy selection task reveals that people avoid empathy for strangers and that people can be motivated to engage empathy (Ferguson et al., 2020). Of note, despite only preliminary reports of the task's reliability (i.e., split-half reliability; Ferguson et al., 2020), individual difference research has already been conducted with the task, perhaps prematurely (Cameron & Inzlicht, 2020; Hu et al., 2020).

The present study

The purpose of the present research is to investigate the reliability of the empathy selection task within a single session and across testing sessions. Prior work has reported split-half (odd-even) reliability estimates ranging from 0.56 to 0.80 (Ferguson et al., 2020), but summary score stability over time has yet to be investigated. Our intention with the present work is to establish the empathy selection task's test-retest reliability within and between testing sessions, and its suitability for use as an individual difference measure.

We recruited participants from an online participant pool (Mechanical Turk) and an undergraduate participant pool, as both commonly used samples in psychological research (Buhrmester et al., 2011; Hauser & Schwarz, 2016). We first assess for rank order stability of scores on the empathy selection task (i.e., proportion of empathy choices) across testing sessions. Next, we examine internal consistency via permutation-based split-half estimation, and task performance variability elicited within and between test sessions. We assess systematic changes at the trial level using generalized multilevel modeling which takes into account participants' individual performance variation.

Method

Participants

Cameron et al. (2019) reported a large meta-analytic empathy avoidance effect (Hedges' $g = -0.64$). A power analysis using G*Power indicated that we could achieve 80% power to detect a medium effect ($d = 0.45$) with as few as 41 participants in our within-subject design. For sample 1 we recruited 60 participants for a two-part study on Mechanical Turk. From this initial sample, 10 participants did not return for part 2 of the study, two participants were removed because they skipped at least one question on the empathy selection task, and one was removed for providing nonsense responses, leaving a final sample of 47 participants ($n = 47$; 18 female, 29 male, $M_{\text{age}} = 35.9$ years, $SD_{\text{age}} = 9.67$). Participants in study 1 were required to correctly answer two comprehension questions before they could move forward with the task.

For sample 2 we recruited 82 undergraduate students for a two-part study at the University of Toronto Scarborough campus. From this initial sample, 14 participants did not return for part 2 of the study, one participant was removed because they skipped at least one question on the empathy selection task, one was removed for providing only nonsensical keyword responses, and 24 were removed because they failed one or more comprehension questions (described in detail below), leaving a final sample of 42 participants ($n = 42$; 34 female, 8 male, $M_{\text{age}} = 19$ years, $SD_{\text{age}} = 2.10$).

Procedure

At time 1, participants in sample 1 completed the empathy selection task followed by the NASA Task Load Index (Hart & Staveland, 1988), the Interpersonal Reactivity Index (Davis, 1980), and a demographic questionnaire. The procedure at time 2 was identical with the exception that we administered the Empathy Index (Jordan et al., 2016) in place of the Interpersonal Reactivity Index. In sample 2, participants completed the empathy selection task followed by the NASA Task Load Index, the Interpersonal Reactivity Index, and a demographic questionnaire measures at both time 1 and time 2. The NASA Task Load Index is an assessment of subjective cognitive effort and was not analyzed in the present work. Time 1 and time 2 were two weeks apart in both studies. In sample 1, all measures were presented using Qualtrics software (Qualtrics, Provo, UT). In sample 2, the empathy selection task was presented using PsychoPy (Peirce et al., 2019) and questionnaires were presented using Qualtrics.

Empathy selection task The version of the empathy selection task used here was identical to the version used in study 3 of Cameron et al. (2019). This version of the task was selected because the empathy and non-empathy deck instructions are very closely matched in terms structure and complexity, and it therefore represents a conservative test of empathy avoidance on the empathy selection task (i.e., the differences between deck activities is intentionally minimized). Participants in both samples were asked to read instructions for the task, and they were informed that they will be quizzed on the instructions. See Appendix 1 for full instructions and see OSF for materials (<https://osf.io/ea3jy/>).

After reading the instructions but prior to beginning the empathy selection task, all participants were required to answer the following two multiple choice questions: “Which of the following is an appropriate response on trials where you are told to be objective?” and “Which of the following is an appropriate response on trials where you are told to be empathic?” Response options for both questions were (a) “Be emotional, and provide keywords describing the internal emotional experiences you are feeling,” (b) “Be objective, and provide keywords describing the facial emotion expression of the person,” and (c) “Be empathic, and provide keywords describing the internal emotional experiences of the person.” The comprehension questions were asked at both time points. In sample 1, participants were not able to move forward with the task until they correctly answered both questions. In sample 2, participants were able to start the task even if they incorrectly answered one or both comprehension questions, and those who failed one or both at either time point were removed prior to analysis.

After the comprehension questions, participants began the trial phase of the empathy selection task. At the beginning of each trial, participants were shown a pair of card decks on the screen in front of them. The deck on the left was labeled “DECK A,” and corresponded to the objective instructions; the deck on the right was labeled “DECK B,” and corresponded to the empathy trials. Participants were instructed to choose a deck, and after a choice was made, they were shown an image of an angry face from the Chicago Face Database (Black and White male and female adults; Ma, Correll, & Wittenbrink, 2015). If participants chose DECK A, they read the following instructions (emphasis in original): “Look at the person in the picture, and try to identify the emotion of this person. Objectively focus on the external facial expression of this person. Please write 3 keywords describing the objective facial expression of this person.” If participants chose DECK B, they were instructed to “Look at the person in the picture, and try to feel what this person feels. Empathically share in the internal emotional experience of this person. Please write 3 keywords describing the subjective emotional experience of this person.”

Participants completed 40 trials. See Fig. 1 for visualization of trial procedure.

Statistical analysis

Behavioral data Statistical analyses were performed in R (R Core Team, 2021). First, empathy choice at each session was estimated by fitting two-level generalized linear models (binomial distribution) with deck choice on each trial nested within participant. Models were estimated using the `glmer()` function within the R package `lme4` (Bates et al., 2015). Estimates of empathy choice were obtained from intercept values (i.e., without predictors in the model). Participant was the only random factor in these models.

Test-retest reliability of summary scores Reliability of empathy choice summary scores (i.e., proportion of empathy deck choice at each session) was assessed with the Pearson's product-moment correlation coefficient (Pearson's r) and the intraclass correlation coefficient (ICC). Pearson's r indicates the strength of the linear relationship between two observations, but it does not take into account systematic error (Rousson et al., 2002). The ICC accounts for both consistency and systematic changes in the group of participants (Vaz et al., 2013). ICCs were calculated using a two-way random effects model for absolute agreement (i.e., ICC [2,1]; Shrout & Fleiss, 1979), which is sensitive to differences between session means. ICC calculations were conducted via the `irr` package in R (Gamer et al., 2019).

Interpretation of reliability metrics is somewhat controversial, and there are no clear guidelines for interpreting reliability values (Crocker & Algina, 1986). It is often suggested that a value of .6 should be considered “good” reliability (e.g., see Cicchetti & Sparrow, 1981; Fleiss, 1981; Landis & Koch, 1977), though a distinction has been made for clinically relevant tasks which suggests a minimum value of .7 and ideal value of .9 (Barch et al., 2008).

As both measurement error and between-participant variability are important for the interpretation of reliability, we also report the standard error of measurement (SEM). A large SEM relative to the between-subject variance contributes to poor reliability (e.g., see Hedge et al., 2018). The SEM is the square root of the error variance term in the ICC calculation and represents the 68% confidence interval around an individual's observed score. We also present reliability metrics for the Interpersonal Reactivity Index (Empathic Concern and Perspective Taking subscales only, Murphy et al., 2020; administered at both time points in sample 2/in-lab only), the Stroop error rate, and the go/no-go commission error rate, so as to compare the empathy selection task to an instrument designed for individual difference research (i.e., the Empathic Concern and Perspective

Taking subscales of the Interpersonal Reactivity Index) and to other commonly used experimental tasks. The Stroop and go/no-go data presented here were collected and made freely available by Hedge et al. (2018). We chose to present the Stroop error rate and the go/no-go commission error rate because these are two commonly used task metrics in individual difference and experimental research, and because they demonstrate relatively weak (Stroop error rate) and strong (go/no-go commission error rate) test-retest reliability. All materials, analysis code, and data are available on the Open Science Framework (<https://osf.io/ea3jy/>).

Trial-level data Internal consistency of empathy avoidance was determined via permutation-based split-half reliability estimates (5000 random splits), obtained separately for each session, using the `splithalf` package for R (version 0.7.1; Parsons, 2020). This approach takes 5000 random (without replacement) split halves of the data, and for each permutation the correlation between halves is calculated, with the Spearman-Brown correction applied (i.e., which takes into account that the estimates are drawn from half the trials that they could have been, Spearman, 1904). However, given that the choice presented at each trial on the empathy selection task is identical, and that the stimuli used in this version of the task were consistent in terms of emotional content (i.e., all faces expressing anger)³, estimations of item consistency are less relevant than they would be for a questionnaire wherein each item is unique and presented only once. Here, we are primarily interested in the empathy choice variability for participants within and between testing sessions (e.g., practice effects; people might remember the format of the task and remember which deck they prefer, and therefore choose it more often in the second session), and whether these outcomes varied by sample (i.e., online vs. in-lab). We fit a three-level generalized multilevel model (binomial distribution) to account for participants' performance variation within and between each testing session. Trial responses were nested within session (i.e., time 1 or time 2), which were nested within participant. Trial number,

³ To assess whether consistency of emotion expression in stimuli impacts internal consistency of the empathy selection task, we computed permutation-based split-half reliability estimates (using 5000 random splits) for versions of the task with multi-emotion stimuli. Specifically, we calculated estimates for Experiments 1a, 1b, 2b, and 3a from Ferguson et al. (2020), because these experiments had 40 trials of the empathy selection task and all included both positive- and negative-valenced images. The data from Ferguson et al. (2020) are also publicly available on OSF (<https://osf.io/qaxgj/>). These calculations suggested Spearman-Brown-corrected reliability estimates from Experiments 1a, 1b, 2b, and 3a of Ferguson et al. (2020) ranged from $r_{SB}=0.85$ to 0.96, which is similar to the estimates reported in the presently reported version of the empathy selection task featuring only angry faces.

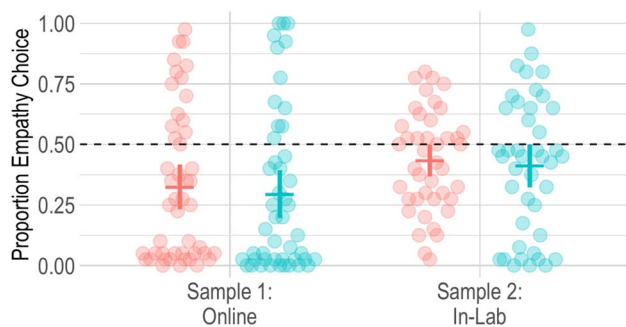


Fig. 2 Beeswarm plot of individual participants' proportion of empathy deck choice on the empathy selection task in sample 1 (online; $n=47$) and sample 2 (in-lab; $n=42$). Horizontal line indicates session average. Vertical line indicates standard error. Dashed line indicates chance level (i.e., 50%) of choice

session, sample, and all interactions were included as fixed effects in this model. Participant and session were included as random effects. A random slope for trial number was included to account for cross-level interactions.

The dependent variable for all multilevel models was trial-by-trial responses on the empathy selection task, and all multilevel models had unstructured covariance matrices. The `glmer()` function relies on an adaptive Gauss–Hermite likelihood approximation to fit the model to the data, and our models were conducted using the Laplace approximation (Liu & Pierce, 1994). The `glmer()` function determines p -values associated with each statistic based on asymptotic Wald tests. Predicted probabilities were calculated using the R package `sjPlot` (Lüdtke, 2020). Effect sizes were calculated via the `r2()` function from the package `performance` (Lüdtke et al., 2019). Overall model effect sizes are presented as *conditional* R^2 , which considers both the fixed and random effects and indicates how much of the model's variance is explained by the complete model (Lüdtke et al., 2019). Fixed effect sizes are presented as *marginal* R^2 .

Results

Behavioral analyses

Summary scores for empathy choice at times 1 and 2 are presented in Fig. 2 and Table 1. As expected, across samples participants avoided the empathy deck at both time 1 (sample 1: $b = -1.30$, $SE = 0.32$, $p < .001$, $R^2_{\text{conditional}} = 0.58$; sample 2: $b = -0.34$, $SE = 0.15$, $p = .028$, $R^2_{\text{conditional}} = 0.21$), and time 2 (sample 1: $b = -1.77$, $SE = 0.46$, $p < .001$, $R^2_{\text{conditional}} = 0.73$; sample 2: $b = -0.69$, $SE = 0.29$, $p = .018$, $R^2_{\text{conditional}} = 0.50$).

Table 1 Proportion of empathy deck choice of the empathy selection task at each time point

Sample	Time 1	Time 2
	Empathy choice [95% CI]	Empathy choice [95% CI]
Study 1 (online)	0.32 [0.23–0.42]	0.29 [0.20–0.39]
Study 2 (in-lab)	0.43 [0.37–0.50]	0.41 [0.32–0.50]

Test-retest reliability

Reliability and agreement metrics for the empathy selection task summary scores (i.e., proportion empathy choice), scores on the Empathic Concern and Perspective Taking subscales of the Interpersonal Reactivity Index, Stroop error cost, and go/no-go commission rate are presented in Table 2. The scatterplots in Fig. 3 demonstrate the size of the SEM in each measure (in original units) relative to the variance in the data. A large SEM relative to the between-person variance in the data contributes to poor reliability (i.e., SEM is in the denominator and between-person variability in the numerator of the ICC calculation). The empathy selection task exceeded a standard of good/substantial reliability (.6) in sample 1, $r = 0.67$ [0.48–0.80], $ICC = 0.67$ [0.48–0.80], $SEM = 0.18$, and sample 2, $r = 0.68$ [0.47–0.81], $ICC = 0.65$ [0.43–0.80], $SEM = 0.15$. The Empathic Concern and Perspective Taking⁴ subscales of the Interpersonal Reactivity Index demonstrated excellent to good reliability (Empathic Concern: $r = 0.82$ [0.69–0.90], $ICC = 0.81$ [0.68–0.89], $SEM = 0.30$; Perspective Taking: $r = 0.78$ [0.63–0.88], $ICC = 0.78$ [0.63–0.88], $SEM = 0.26$). The two task indices from Hedge et al. (2018) demonstrated moderate (Stroop error cost, $r = 0.45$ [0.22–0.63], $ICC = 0.44$ [0.20–0.63], $SEM = 2.45$) and good/substantial (go/no-go commission rate, $r = 0.79$ [0.67–0.87], $ICC = 0.76$ [0.60–0.86], $SEM = 6.46$) reliability.

The relationship between the ICC and the sources of variance across measures is shown in Fig. 4, which plots the three components of variance from which the ICC is calculated. The size of the bar is normalized for the total amount of variance in each measure and is decomposed into variance accounted for by differences between individuals (grey), by differences between sessions (white), and error variance (black). Correlational research (and the ICC) relies on the proportion of variance accounted for by differences between individuals. The Stroop error cost demonstrated the

⁴ One participant scored more than three median absolute deviations (MAD) away from the median on the Perspective Taking subscale and was removed prior to analysis. Including this participant does not meaningfully change the results presented here.

Table 2 Test-retest reliability metrics of the empathy selection task, Interpersonal Reactivity Index, Stroop error cost, and go/no-go commission rate. ICCs are typically interpreted as excellent (>.8), good/

substantial (>.6), and moderate (>.4) (Cicchetti & Sparrow, 1981; Fleiss, 1981; Landis & Koch, 1977)

Measure	Sample	<i>r</i> [<i>CI</i>]	<i>ICC</i> [<i>CI</i>]	<i>SEM</i>
Empathy selection task	Sample 1 (online)	0.67 [0.48–0.80]	0.67 [0.48–0.80]	0.19
	Sample 2 (in-lab)	0.68 [0.47–0.81]	0.65 [0.43–0.80]	0.15
Empathic Concern (IRI)	Sample 2 (in-lab)	0.82 [0.69–0.90]	0.81 [0.68–0.89]	0.30
Perspective Taking (IRI)	Sample 2 (in-lab)	0.78 [0.63–0.88]	0.78 [0.63–0.88]	0.26
Stroop error cost	Hedge et al. (study 2)	0.45 [0.22–0.63]	0.44 [0.20–0.63]	2.45
GNG commission rate	Hedge et al. (study 2)	0.79 [0.67–0.87]	0.76 [0.60–0.86]	6.46

Sample 1 $n = 47$, sample 2 $n = 42$, $r =$ Pearson's r ; $ICC =$ intraclass correlation coefficient; $CI = 95\%$ confidence interval; $SEM =$ standard error of measurement; IRI = Interpersonal Reactivity Index; GNG = go/no-go. SEMs are in the original units

lowest ICC which may be in part because it is a calculated difference score (i.e., errors made on congruent trials minus those made on incongruent trials). By design, within-subject indices like difference scores reduce the between participant variance relative to error variance. This is desirable for researchers interested in within-person variation in task performance but problematic for researchers interested in between-person differences. Here, we find that the empathy selection task demonstrates a relatively healthy (“good/substantial”) proportion of variance accounted for by individuals, approximately on par with or superior to the two other experimental tasks discussed here, but less than the Interpersonal Reactivity Index.

Trial-level analyses

Internal consistency of empathy choice was estimated using a permutation-based split-half approach (Parsons, 2020) with 5000 random splits. In sample 1, the Spearman-Brown-corrected split-half internal consistency (associated 95% confidence intervals presented in brackets) of empathy deck choice on the empathy selection task was $r_{SB} = 0.97$ [0.95–0.98] at time 1 and $r_{SB} = 0.98$ [0.96–0.99] at time 2. In sample 2, the Spearman-Brown-corrected split-half internal consistency of empathy choice was $r_{SB} = 0.88$ [0.82–0.93] at time 1 and $r_{SB} = 0.95$ [0.92–0.97] at time 2.

The analysis evaluating within- and between-session effects demonstrated a significant main effect of sample on empathy choice, $b = 1.28$, $SE = 0.57$, $p = .026$, such that participants in sample 2 (in lab; proportion empathy choice = 0.42 [0.35, 0.49]) were more likely to choose the empathy deck than those in sample 1 (online; proportion empathy choice = 0.31 [0.22, 0.40]). Replicating Cameron et al. (2019), there was a main effect of trial number, $b = -0.03$, $SE = 0.01$, $p = .013$, such that empathy deck choice decreased as participants progressed through the task. There was not a main effect of session ($p = 0.133$), and there were no interactions between any of the fixed effects ($ps > .081$).

The combined fixed effect size of trial number, session, and sample in this model was $R^2_{\text{marginal}} = 0.048$. Overall model effect size, including both fixed and random effects, was $R^2_{\text{conditional}} = 0.571$. See Table 3 and Fig. 5 for model details.

Discussion

We investigated the reliability of a novel task of empathy behavior, the empathy selection task, to identify systematic performance changes across and within testing sessions and commonly used research samples (i.e., online, undergraduates). Summary scores on the empathy selection task demonstrated good/substantial test-retest reliability (ICCs = .65 and .67), though it was considerably lower than that demonstrated by the Empathic Concern (ICC = .81) and Perspective Taking (ICC = .78) scales of the Interpersonal Reactivity Index, and short of the conservative suggestion for cognitive and clinical measures (.70; Barch et al., 2008).

Recently, Hedge et al. (2018) documented test-retest reliabilities across a range of cognitive measures (e.g., Stroop, stop-signal task) and found that the majority of metrics had ICC values below .70, and many had values much lower than that. While the go/no-go commission error rate performed fairly well in terms of reliability (ICC = .76), the Stroop error rate was far less reliable (ICC = .44). Together, our results suggest that the summary scores derived from the empathy selection task (i.e., proportion empathy deck choice) may be as appropriate for use in individual difference research as some commonly used experimental tasks like the go/no-go commission error rate, but that it is likely less suited to such research than self-report questionnaires such as the Interpersonal Reactivity Index.

The mixed effects analysis, used here to better understand the temporal stability of the empathy selection task both within and between sessions across samples, showed that deck preference on the empathy selection task was stable across testing sessions but that empathy deck choice steadily

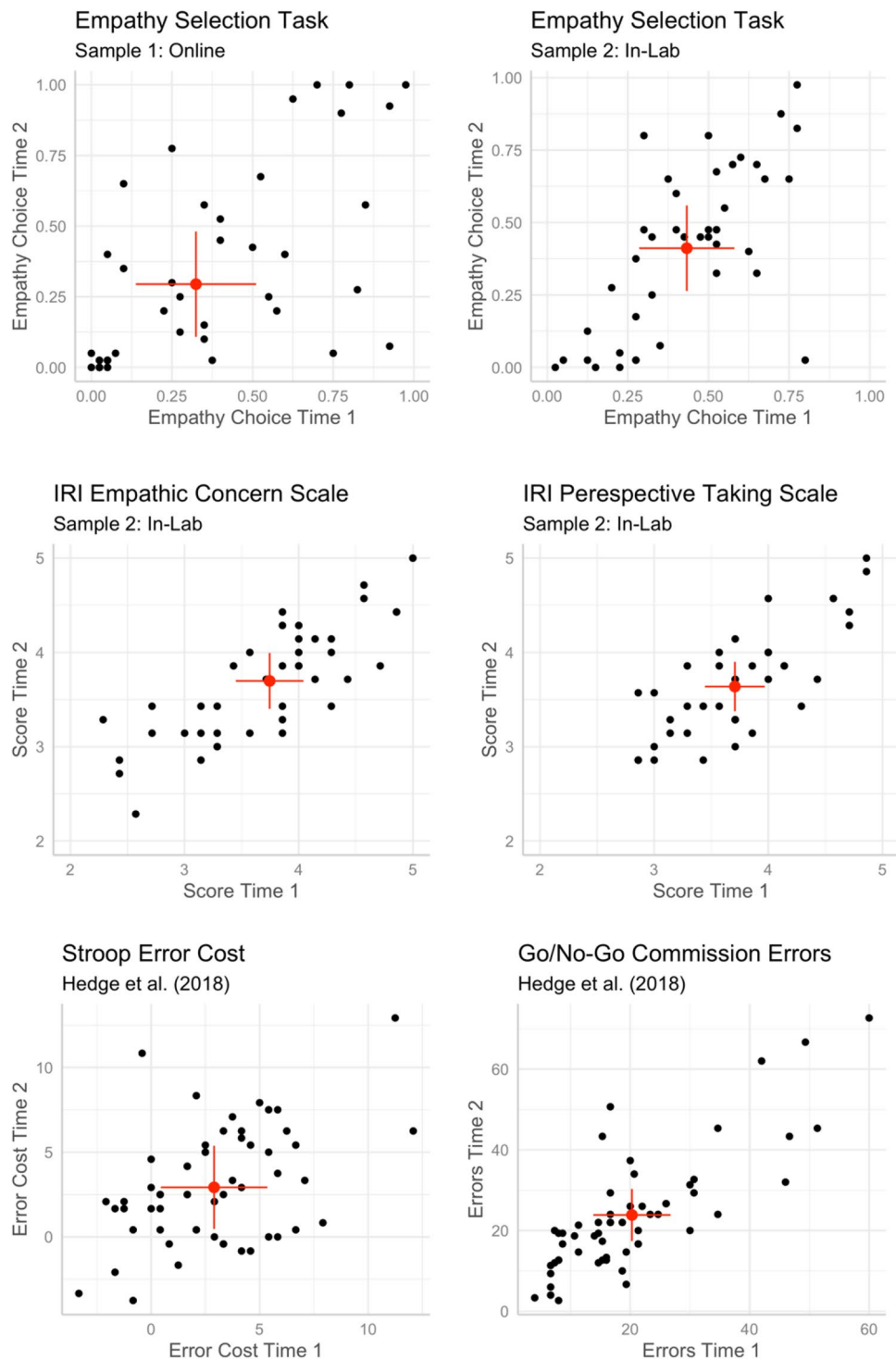


Fig. 3 Reliability of summary scores on the empathy selection task (sample 1, $n=47$; sample 2, $n=42$), scores on the Empathic Concern and Perspective Taking scales from the Interpersonal Reactivity Index (sample 2, $n=42$), Stroop error cost (from study 2 in Hedge et al., 2018, $n=56$), and go/no-go commission rate (from study 2 in

Hedge et al., 2018, $n=57$). Red markers indicate mean group performance at time 1 and time 2. Error bars show ± 1 standard error of measurement (SEM). The SEM is the square root of the error variance term calculated from the intraclass correlation, and can be interpreted as the 68% confidence interval for an individual's data point

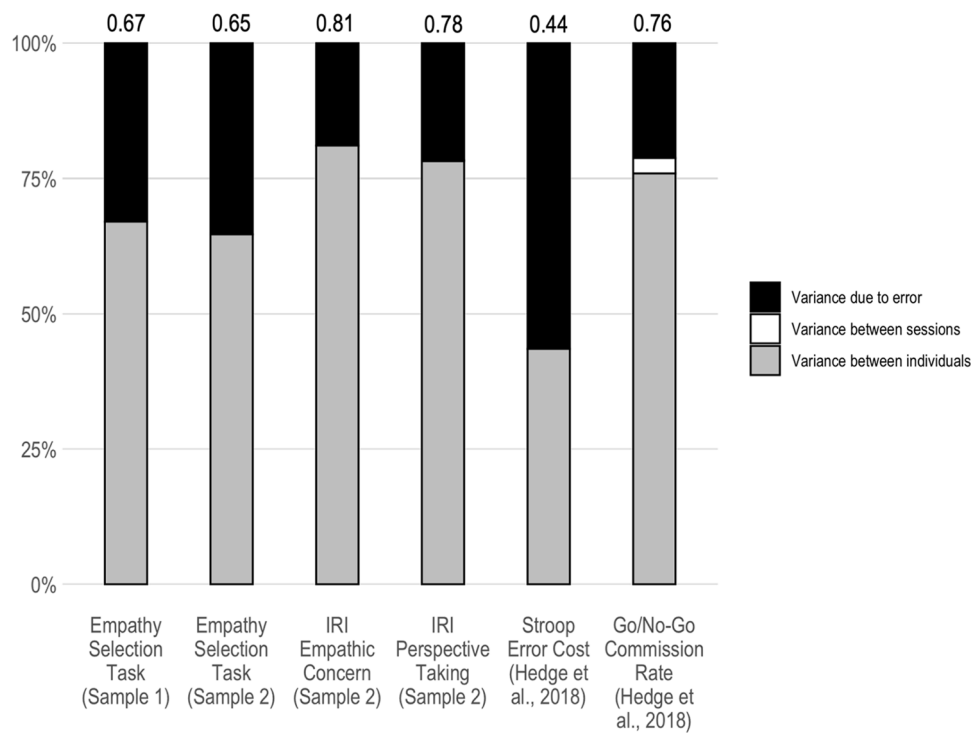


Fig. 4 Relative size of variance components for the empathy selection task (sample 1, $n=47$, sample 2, $n=42$), scores on the Empathic Concern and Perspective Taking scales from the Interpersonal Reactivity Index (sample 2, $n=42$), Stroop error cost (from study 2 in Hedge et al., 2018, $n=56$), and go/no-go commission rate (from

study 2 in Hedge et al., 2018, $n=57$). The intraclass correlation (ICC) is printed above the bar. The size of the bar is normalized for the total amount of variance in the measure and subdivided into variance accounted for by differences between individuals (gray), by differences between sessions (white), and error variance (black)

declined within a single testing session. That is, we did not find evidence of practice effects (e.g., remembering one’s preferred deck from time 1 and choosing it more often at time 2) across sessions, but empathy deck choice declined as trial number increased. This is consistent with previous research demonstrating a small negative relation between empathy choice and time on task (odds ratio = .98 in Cameron et al., 2019) and with other paradigms which measure effort avoidance (e.g., the demand selection task, Kool et al., 2010).

Table 3 Results from the generalized mixed effects analysis of empathy choice

Predictors	Odds ratio	CI	<i>p</i>
(Intercept)	0.27	0.16–0.48	<.001
Trial no.	0.97	0.95–0.99	0.013
Session	0.73	0.48–1.10	0.133
Sample	3.60	1.17–11.09	0.026
Trial no. × session	0.99	0.97–1.01	0.324
Trial no. × sample	1.04	0.99–1.09	0.081
Session × sample	1.02	0.44–2.34	0.97
Trial no. × session × sample	1.02	0.99–1.06	0.231

Note: Bolded values indicate $p < .05$

We found empathy avoidance in both samples, but individuals in the online sample were less likely to choose the empathy deck overall than those in the lab sample. Understanding the theoretical basis of the empathy selection task is crucial to one’s interpretation of this finding. One explanation is that online participants are simply less empathic than the lab participants (though follow-up analyses demonstrated no differences between the samples in IRI Empathic Concern or IRI Perspective Taking scores). A more likely explanation is that individuals who are completing the task through Mechanical Turk are particularly motivated to expend as little effort as possible in the service of completing as many paid assignments as fast as possible. Thus, it may be expected that online samples would tend toward the lower-effort task (in this case, the non-empathy deck), in the service of efficiency and to avoid fatigue. This finding highlights the importance of understanding the theoretical basis of the empathy selection task before putting it to use. Here, in the absence of this understanding, differences in empathy choice across samples might be incorrectly attributed to true differences in empathy avoidance rather than individual differences in effort-sensitivity which might be unrelated to empathy entirely.

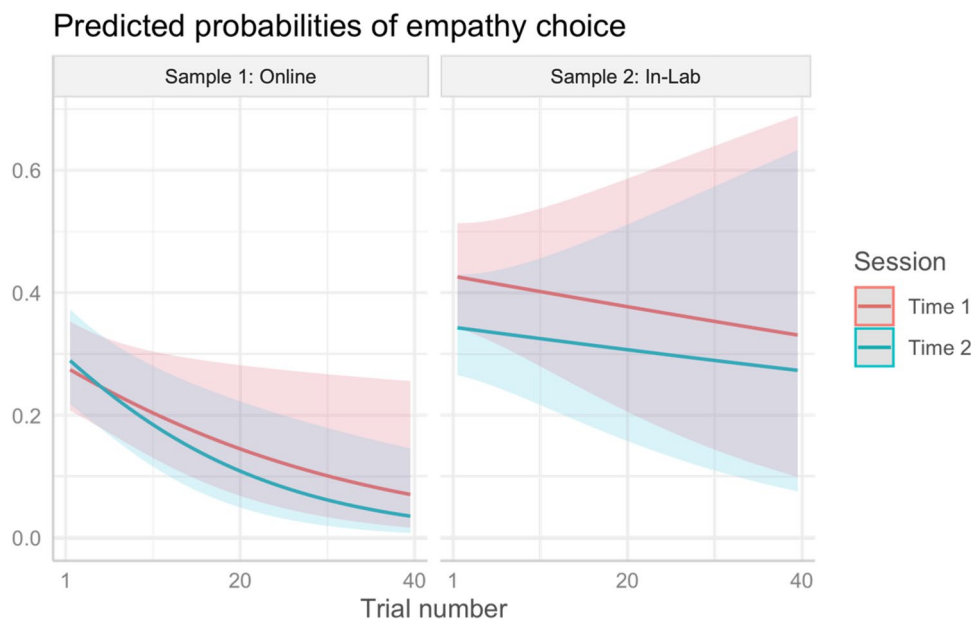


Fig. 5 Predicted probabilities of empathy choice as a function of fixed effects of time on task (i.e., trial number), session (i.e., time 1 or time 2), and sample (i.e., online or in-lab). Dotted line denotes chance level (i.e., 50%) probability of choosing the empathy deck

There were a relatively large number of participants removed from sample 2 following incorrect response on one or both comprehension questions. It is possible that the instructions provided were long enough that participants skimmed instead of reading thoroughly, despite additional instruction to expect comprehension checks (i.e., “you will be quizzed”). However, the comprehension questions were quite simple, and used identical language in the response options (e.g., “Which of the following is an appropriate response on trials where you are told to be empathic?”, correct answer: “Be empathic,...”; “Which of the following is an appropriate response on trials where you are told to be objective?”, correct answer: “Be objective,...”), which suggests that people were not reading the instructions at all rather than not understanding them. This is somewhat welcome news, as it implies that the relatively high comprehension failure rate observed in sample 2 is more likely due to a lack of attention than challenging or unwieldy instructions. Nonetheless, it will be important for future researchers interested in using the empathy selection task to assess for inattention and random responding. This is perhaps especially true for those interested in collecting data from online samples, as concerns about non-English-speaking workers and/or bots on Mechanical Turk have increased in recent years (e.g., see Chmielewski & Kucker, 2020).

The sample 1/online data presented here were collected in June 2018, just prior to the observed shift in Mechanical Turk data quality (e.g., as described by Chmielewski & Kucker, 2020). In our subsequent work using the empathy selection task on Mechanical Turk, we have noticed

a steep decline in data quality (e.g., see study 3c from Ferguson et al., 2020). Best practices for screening non-English-speaking workers and/or bots should be followed (see Chmielewski & Kucker, 2020 for excellent examples of response validity indicators) when using the empathy selection task online. Similarly, establishing predefined trial-level response criteria can also assist with data quality. For example, since the instructions for each trial were present on every trial (“provide three keywords...”), a failure to provide any of the information requested indicates a significant lack of understanding of task instructions and/or significant task disengagement. Only two participants met these criteria in the present work (one in each sample) and were removed for this reason.

Limitations Here, we tested a version of the empathy selection task which has carefully matched empathy and non-empathy tasks (i.e., the only difference in the decks was the requirement to empathize in the empathy deck), and which uses a limited set of emotion stimuli (i.e., angry faces from the Chicago Face Database). It is possible that different iterations of the empathy selection task may be more or less reliable than the version tested here. Cameron et al. (2019) tested multiple versions of the empathy selection task, including varied deck labels, position, images displayed, and response style (e.g., written, binary). While behavioral outcomes across these versions was consistent in that people tended to avoid the empathy deck, it is possible that different versions will have different between-person, between-session, and error variance components, making

them more or less reliable and/or suitable for individual difference research. For example, it is likely that more trials of the task will increase test-retest reliability—from a classical test theory perspective, increasing a test's length will typically reduce measurement error (Kuder & Richardson, 1937; see also Rouder & Haaf, 2019).

Similarly, a limitation in the present work is that the version of the empathy selection task used here presented only images of angry faces. It is possible that using a wider range of stimuli, particularly in the absence of additional trials, may prolong the learning phase of the task which may have downstream consequences to its reliability. That is, it is possible that image valence is one of several parameters assessed by participants when establishing a preference between the decks. Increasing the variation in image valence may therefore extend the time taken to establish such a preference. While previous work has demonstrated that individuals equally avoid empathy for positive and negative emotions on the empathy selection task (see studies 4–6 in Cameron et al., 2019), the extent to which these differences impact task reliability is an open question.

A limitation that our study shares with several other related studies is limited sample size. The differences in empathy choice across samples suggests the possibility of measurement non-invariance across samples, and future work might examine measurement invariance of the empathy selection task using a larger sample than was available in the present work. Similarly, limited sample size can affect the stability of the test-retest metrics. While previous work with the empathy selection task has reported split-half reliability (ranging from 0.56 to 0.80; Cameron et al., 2019; Ferguson et al., 2020), our results should be considered as the first test-retest reliability estimates for the empathy selection task, with more work needed to refine and maximize its capacity for use in individual difference research.

Conclusion

The empathy selection task is a novel behavioral measure of empathy avoidance, and there are likely to be individual-level factors that predict someone's willingness to engage in empathy on this task. This is certainly an area of interest in psychological science, but we must first assess the suitability of our assessments for use in correlational research. We find that the empathy selection task demonstrates a good/substantial ability to consistently rank individuals, and that it is temporally stable within and between testing sessions. We therefore conclude that the empathy selection task can be used for individual difference research.

Appendix 1 – Empathy selection task instructions

After completing a consent form, participants read the following instructions before beginning the empathy selection task:

Please read these instructions carefully (you will be quizzed!)

In this task, you will complete a series of trials. On each trial, you will see two decks of cards. You should choose between these decks. Once you choose a deck, you will then see an image of a person. These are actors that we've asked to look certain ways and express certain emotions. Depending on which deck you have chosen, you will then be given one of two possible sets of instructions.

On some trials, you will be told to be objective and focus on identifying the emotional expression of the person in the image. When completing this kind of trial, try to be objective. To be objective, try to focus on which emotions the person's facial expression most closely resembles. On these trials, please provide three keywords to describe the facial expression of the person (Example: "sad, hurt, confused" or "happy, pleased, interested"). It is okay to use the same keywords multiple times, just make sure that you are describing the emotional expression of the person in the image.

On other trials, you will be told to have empathy and share in the emotional experience of the person in the image. When completing this kind of trial, try to feel empathy. To be empathic, try to share in the internal emotional experience of the person. On these trials, please provide three keywords to describe the emotional experience of this person (Example: "sad, hurt, confused" or "happy, pleased, interested"). It is okay to use the same keyword multiple times, just make sure you are describing the feelings and experiences of the person in the image.

You are free to choose from either deck on any trial, and should feel free to move from one deck to the other whenever you choose. If one deck begins to seem preferable, feel free to choose that deck more often. Overall, this task will take the same amount of time regardless of which deck you choose.

Press continue when you are ready to begin.

Acknowledgements This work was supported by a Social Sciences and Humanities Research Council of Canada grant (#435-2019-0144) and a Natural Sciences and Engineering Research Council of Canada grant (RGPIN-2019-05280) held by the last author.

References

- Andreoni, J., Rao, J. M., & Trachtman, H. (2017). Avoiding the ask: A field experiment on altruism, empathy, and charitable giving. *Journal of Political Economy*, *125*(3), 625–653.
- Barch, D. M., Carter, C. S., Combs, C. E., et al. (2008). Measurement issues in the use of cognitive neuroscience tasks in drug development for impaired cognition in schizophrenia: A report of the second consensus building conference of the CNTRICS initiative. *Schizophrenia Bulletin*, *34*, 613–618. <https://doi.org/10.1093/schbul/sbn037>
- Baron-Cohen, S., Jolliffe, T., Mortimore, C., & Robertson, M. (1997). Another advanced test of theory of mind: Evidence from very high functioning adults with autism or Asperger syndrome. *Journal of Child Psychology and Psychiatry*, *38*(7), 813–822.
- Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., ... & Rcpp, L. (2015). Package 'lme4'. *Convergence*, *12*(1), 2.
- Batson, C. D. (2009). *These things called empathy: Eight related but distinct phenomena*. In J. Decety & W. Ickes (Eds.), *Social neuroscience. The social neuroscience of empathy* (pp. 3–15). MIT Press. <https://doi.org/10.7551/mitpress/9780262012973.003.0002>
- Batson, C. D., Klein, T. R., Highberger, L., & Shaw, L. L. (1995). Immorality from empathy-induced altruism: When compassion and justice conflict. *Journal of Personality and Social Psychology*, *68*(6), 1042.
- Bernhardt, B. C., & Singer, T. (2012). The neural basis of empathy. *Annual Review of Neuroscience*, *35*, 1–23.
- Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 3–5.
- Cameron, C. D., & Inzlicht, M. (2020). Empathy choice in physicians and non-physicians. *British Journal of Social Psychology*, *59*(3), 715–732.
- Cameron, C. D., & Payne, B. K. (2011). Escaping affect: How motivated emotion regulation creates insensitivity to mass suffering. *Journal of Personality and Social Psychology*, *100*(1), 1–15.
- Cameron, C. D., Harris, L. T., & Payne, B. K. (2016). The emotional cost of humanity: Anticipated exhaustion motivates dehumanization of stigmatized targets. *Social Psychological and Personality Science*, *7*(2), 105–112.
- Cameron, C. D., Hutcherson, C. A., Ferguson, A. M., Scheffer, J. A., Hadjiandreou, E., & Inzlicht, M. (2019). Empathy is hard work: People choose to avoid empathy because of its cognitive costs. *Journal of Experimental Psychology: General*, *148*(6), 962.
- Chmielewski, M., & Kucker, S. C. (2020). An MTurk Crisis? Shifts in Data Quality and the Impact on Study Results. *Social Psychological and Personality Science*, *11*(4), 464–473.
- Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *American Journal of Mental Deficiency*, *86*(2), 127–137.
- Cosley, B. J., McCoy, S. K., Saslow, L. R., & Epel, E. S. (2010). Is compassion for others stress buffering? Consequences of compassion and social support for physiological reactivity to stress. *Journal of Experimental Social Psychology*, *46*(5), 816–823.
- Cramer, D., & Jowett, S. (2010). Perceived empathy, accurate empathy and relationship satisfaction in heterosexual couples. *Journal of Social and Personal Relationships*, *27*(3), 327–349. <https://doi.org/10.1177/0265407509348384>
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston.
- Dang, J., King, K. M., & Inzlicht, M. (2020). Why are self-report and behavioral measures weakly correlated? *Trends in Cognitive Sciences*, *24*(4), 267–269.
- Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology*, *10*, 85.
- Davis, M. H., Luce, C., & Kraus, S. J. (1994). The heritability of characteristics associated with dispositional empathy. *Journal of Personality*, *62*(3), 369–391.
- Decety, J., & Jackson, P. L. (2004). The functional architecture of human empathy. *Behavioral and Cognitive Neuroscience Reviews*, *3*(2), 71–100.
- Depow, G. J., Francis, Z., & Inzlicht, M. (2021). The experience of empathy in everyday life. *Psychological Science*, *32*(8), 1198–1213. <https://doi.org/10.1177/0956797621995202>
- Dunn, T. L., Inzlicht, M., & Risko, E. F. (2019). Anticipating cognitive effort: roles of perceived error-likelihood and time demands. *Psychological Research*, *83*(5), 1033–1056.
- Ferguson, A. M., Cameron, D., & Inzlicht, M. (2020). Motivational effects on empathic choices. *Journal of Experimental Social Psychology*, *90*, 104010.
- Ferguson, A. M., Cameron, C. D., & Inzlicht, M. (2021). When does empathy feel good? *Current Opinion in Behavioral Sciences*, *39*, 125–129.
- Fleiss, J. L. (1981). Balanced incomplete block designs for inter-rater reliability studies. *Applied Psychological Measurement*, *5*(1), 105–112.
- Gamer, M., Lemon, J., Singh, I.F.P. (2019). irr: Various Coefficients of Interrater Reliability and Agreement. R package version 0.84.1. <https://CRAN.R-project.org/package=irr>
- Gross, J. J. (2013). Emotion regulation: taking stock and moving forward. *Emotion*, *13*(3), 359–365.
- Haaf, J. M., & Rouder, J. N. (2019). Some do and some don't? Accounting for variability of individual difference structures. *Psychonomic Bulletin & Review*, *26*(3), 772–789.
- Hall, P. A. (2012). Executive control resources and frequency of fatty food consumption: Findings from an age-stratified community sample. *Health Psychology*, *31*(2), 235.
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology*, *52*, 139–183.
- Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*(1), 400–407.
- Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186.
- Hu, T., Zheng, X., & Huang, M. (2020). Absence and presence of human interaction: the relationship between loneliness and empathy. *Frontiers in Psychology*, *11*, 768.
- Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, *100*(3), 828.
- Jordan, M. R., Amir, D., & Bloom, P. (2016). Are empathy and concern psychologically distinct? *Emotion*, *16*(8), 1107.
- Keyes, C., & Gazzola, V. (2014). Dissociating the ability and propensity for empathy. *Trends in Cognitive Sciences*, *18*(4), 163–166. <https://doi.org/10.1016/j.tics.2013.12.011>
- Kool, W., McGuire, J. T., Rosen, Z. B., & Botvinick, M. M. (2010). Decision making and the avoidance of cognitive demand. *Journal of Experimental Psychology: General*, *139*(4), 665.
- Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*(3), 151–160.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.
- Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, *46*(3), 551–556.

- Liu, Q., & Pierce, D. (1994). A note on Gauss—Hermite quadrature. *Biometrika*, 81(3), 624–629. <https://doi.org/10.1093/biomet/81.3.624>
- Lüdtke, D. (2020). *sjPlot: Data Visualization for Statistics in Social Science*. R package version 2.8.6, <https://CRAN.R-project.org/package=sjPlot>
- Lüdtke, D., Waggoner, P. D., & Makowski, D. (2019). Insight: A unified interface to access information from model objects in R. *Journal of Open Source Software*, 4(38), 1412.
- Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, 47(4), 1122–1135.
- MacLeod, C. M. (1991). Half a century of research on the Stroop effect: an integrative review. *Psychological Bulletin*, 109(2), 163.
- Murphy, B. A., & Lilienfeld, S. O. (2019). Are self-report cognitive empathy ratings valid proxies for cognitive empathy ability? Negligible meta-analytic relations with behavioral task performance. *Psychological Assessment*, 31(8), 1062.
- Murphy, B. A., Costello, T. H., Watts, A. L., Cheong, Y. F., Berg, J. M., & Lilienfeld, S. O. (2020). Strengths and weaknesses of two empathy measures: A comparison of the measurement precision, construct validity, and incremental validity of two multidimensional indices. *Assessment*, 27(2), 246–260.
- Neumann, D. L., Chan, R. C., Boyle, G. J., Wang, Y., & Westbury, H. R. (2015). Measures of empathy: Self-report, behavioral, and neuroscientific approaches. In G. J. Boyle, D. H. Saklofske, & G. Matthews (Eds.), *Measures of personality and social psychological constructs* (pp. 257–289). Academic Press.
- Parsons, S. (2020). *Splithalf: Robust estimates of split half reliability*. <https://doi.org/10.6084/m9.figshare.5559175.v5>
- Pavey, L., Greitemeyer, T., & Sparks, P. (2012). “I help because I want to, not because you tell me to” empathy increases autonomously motivated helping. *Personality and Social Psychology Bulletin*, 38(5), 681–689.
- Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. (2019). PsychoPy2: experiments in behavior made easy. *Behavior Research Methods*, 51(1), 195–203. <https://doi.org/10.3758/s13428-018-01193-y>
- Pierce, J. R., Kilduff, G. J., Galinsky, A. D., & Sivanathan, N. (2013). From glue to gasoline: How competition turns perspective takers unethical. *Psychological Science*, 24(10), 1986–1994.
- Preston, S. D., & De Waal, F. B. (2002). Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences*, 25(1), 1–20.
- R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rey-Mermet, A., Gade, M., & Oberauer, K. (2018). Should we stop thinking about inhibition? Searching for individual and age differences in inhibition ability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(4), 501.
- Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review*, 26(2), 452–467.
- Rousson, V., Gasser, T., & Seifert, B. (2002). Assessing intrarater, interrater and test–retest reliability of continuous measurements. *Statistics in Medicine*, 21(22), 3431–3446.
- Scheffer, J. A., Cameron, C. D., & Inzlicht, M. (2022). Caring is costly: People avoid the cognitive work of compassion. *Journal of Experimental Psychology: General*, 151(1), 172–196.
- Shaw, L. L., Batson, C. D., & Todd, R. M. (1994). Empathy avoidance: Forestalling feeling for another in order to escape the motivational consequences. *Journal of Personality and Social Psychology*, 67(5), 879–887. <https://doi.org/10.1037/0022-3514.67.5.879>
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420.
- Spearman, C. (1904). ‘General intelligence,’ objectively determined and measured. *The American Journal of Psychology*, 15(2), 201–293.
- Van Lange, P. A. (2008). Does empathy trigger only altruistic motivation? How about selflessness or justice? *Emotion*, 8(6), 766.
- Vaz, S., Falkmer, T., Passmore, A. E., Parsons, R., & Andreou, P. (2013). The case for using the repeatability coefficient when calculating test–retest reliability. *PLoS One*, 8(9), e73990.
- Zaki, J. (2014). Empathy: A motivated account. *Psychological Bulletin*, 140(6), 1608–1647. <https://doi.org/10.1037/a0037679>

Open practices statement The data and materials for all studies are available at <https://osf.io/ea3jy/>

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.