Article

Assessment
1–15

DOI: 10.1177/1073191118777636
journals.sagepub.com/home/asm
SAGE

Strengths and Weaknesses of Two Empathy Measures: A Comparison of the Measurement Precision, Construct Validity, and Incremental Validity of Two Multidimensional Indices

Brett A. Murphy[1], Thomas H. Costello[1], Ashley L. Watts[1], Yuk Fai Cheong[1], Joanna M. Berg[1], and Scott O. Lilienfeld[1]

Abstract

The quality of empathy research, and clinical assessment, hinges on the validity and proper interpretation of the measures used to assess the construct. This study investigates, in an online sample of 401 adult community participants, the construct validity of the Affective and Cognitive Measure of Empathy (ACME) relative to that of the Interpersonal Reactivity Index (IRI), the most widely used multidimensional empathy research measure. We investigated the factor structures of both measures, as well as their measurement precision across varying trait levels. We also examined them both in relation to convergent and discriminant criteria, including broadband personality dimensions, general emotionality, personality disorder features, and interpersonal malignancy. Our findings suggest that the ACME possesses incremental validity beyond the IRI for most constructs related to interpersonal malignancy. Our results further indicate that the IRI Personal Distress scale is severely deficient in construct validity, raising serious concerns regarding past findings that have included it when computing total empathy scores. Finally, our results indicate that both questionnaires display poor measurement precision at high trait levels, emphasizing the need for future researchers to develop indices that can reliably measure high levels of empathy.

Keywords

empathy, construct validity, incremental validity, psychopathy, personality

In the modern world of the Internet, cable television, and social media, we can now witness others' suffering on a global scale. Perhaps in part as a consequence, empathy has become an increasingly popular topic in the public eye. It is difficult to escape mentions of empathy in political speeches, motivational lectures, religious sermons, and popular psychology articles, among others. Furthermore, empathy has long been a pivotal concept in clinical psychology theory and practice, at least from the time of Carl Rogers (1958). There is substantial debate, however, among researchers and theorists regarding the definition of empathy, whether and how to parse it into subcomponents, and how best to measure it in research and clinical practice. The differences of opinion regarding these definitional and measurement issues have recently spilled over into debates about whether empathy is psychologically beneficial (Baron-Cohen, 2016) or, instead, is actually harmful (Bloom, 2016, 2017).

To provide valuable information relevant to these questions, we aimed to compare the evidence of construct validity of two self-report measures of empathy, one widely used and the other a promising new questionnaire, namely, the Interpersonal Reactivity Index (IRI; Davis, 1983) and the Affective and Cognitive Measure of Empathy (ACME; Vachon & Lynam, 2016), respectively. To do so, we first used confirmatory factor analysis (CFA) and exploratory structural equation modeling (ESEM) methods to examine the tenability of the factor structures of each measure as they are typically adopted in the literature. Second, we examined the measurement precision of these measures using item response theory (IRT) techniques, to compare their respective abilities to reliably detect empathic traits at low and high levels of their latent traits.

Third, we examined these two measures' relations with a broad swath of theoretically relevant external criteria, including general and potentially maladaptive personality traits, with a focus on those associated with interpersonal

[1]Emory University, Atlanta, GA, USA

Corresponding Author:
Brett Murphy, Department of Psychology, Emory University, 36 Eagle Row, Room 270, Decatur, GA 30322, USA.
Email: bmurphy.psych@gmail.com

malignancy (e.g., coldheartedness, meanness). Moreover, we examined these measures' relations to broader emotionality, particularly negative emotionality (NE), a pervasive dimension of distress and emotional maladjustment that courses through most indices of psychopathology (Watson & Clark, 1984). Fourth and finally, in addition to elucidating each empathy measure's nomological networks, we compared the two empathy measures' incremental contributions above and beyond one another in statistically predicting traits highly conceptually associated with interpersonal malignancy.

The Heterogeneity of Empathy

Theorists have posited a wide range of subdimensions (or subtypes) within the empathy construct, including but not limited to effortful perspective-taking, emotion recognition, affective contagion, prosocial motivation, distinctions between self and other, compassion, and aversion to harming others. The closest approximation to a contemporary research vernacular is the separation of empathy into "affective" and "cognitive" components (e.g., Davis, 1983; Shamay-Tsoory, Aharon-Peretz, & Perry, 2009). In this familiar distinction, affective empathy comprises emotional resonance with or caring for the feelings of others. Cognitive empathy, in contrast, comprises the capacity to understand and predict the thoughts and feelings of others.

As noted briefly earlier, the Interpersonal Reactivity Index (IRI; Davis, 1983) is the most widely employed multidimensional research questionnaire of empathy, cited nearly 6,000 times at the time of this writing (according to *Google Scholar*). Aside from the IRI, only four other multidimensional empathy measures intending to encompass cognitive *and* affective empathy have received significant construct validation: the Empathy Quotient (EQ; Baron-Cohen & Wheelwright, 2004), typically analyzed only as a unidimensional global empathy scale; the Basic Empathy Scale (BES; Jolliffe & Farrington, 2006); the Questionnaire of Cognitive and Affective Empathy (QCAE; Reniers, Corcoran, Drake, Shryane, & Völlm, 2011); and, most recently, the ACME. In particular, the ACME attempts to expand the content domain of empathy. Provisional evidence indicates it may be superior to its predecessors (e.g., IRI, BES) in construct validity (Vachon & Lynam, 2016).

The aforementioned measures differ substantially in their conceptualizations of empathy and content coverage. For instance, the BES and the QCAE differ from the other three measures in that they aim to exclude compassion, empathic concern, and other ostensibly prosocial emotional responses from the "affective empathy" construct. This exclusion reflects a major conceptual disagreement regarding the construct of empathy. Some theorists have argued that both the empathizing perceiver and the target individual must be in the *same* emotional state for affective empathy to be present (e.g., Bird & Viding, 2014; Bloom, 2016). More specifically, a number of researchers (cf. Decety & Michalska, 2010) contend that we are "sympathizing" if we feel compassion for someone who is feeling depressed, whereas we are "empathizing" if we, too, feel depressed. Defined narrowly as feeling the same emotion that another is feeling, empathy (a) is likely to be less beneficial than is other-oriented concern (Bloom, 2017), (b) would seem to relate conceptually to heightened negative affect and/or emotional dysregulation, or (c) both.

Other scholars, though, have argued that the perceiver's emotional resonance needs only to be interpersonally "appropriate" (Baron-Cohen, 2016; Baron-Cohen & Wheelwright, 2004) to constitute empathy. This alternative conceptualization does not require that one's emotional state be isomorphic with the emotional state of the other. The IRI, ACME, and EQ follow this broader conceptual perspective, including caring emotional response rather than only isomorphic emotional contagion. This latter definition involves a different perspective on the meaning of emotional resonance, which can be rooted in evolved social functionality (cf., Keltner & Haidt, 1999; Keltner & Kring, 1998). At the same time, though, it may beg the question of what constitutes an interpersonally "appropriate" emotional response.

The IRI Scales

As we have alluded, the heterogeneity of the conceptualization of empathy is reflected both across and within measures. For instance, the IRI contains four scales: Empathic Concern (EC), Perspective Taking (PT), Fantasy (FN), and Personal Distress (PD), each of which aims to reflect differing subdimensions of the broader empathy construct. The EC scale is intended to assess other-oriented feelings of sympathy and concern for others (Davis, 1983), and is largely considered an index of affective empathy. The PT scale is intended to index readiness to adopt another person's perspective, but not necessarily the accuracy of such perspective-taking. Although frequently relied on as a measure of cognitive empathy, this practice is suspect given that the PT scale items' content appears to comprise empathic motivation and agreeableness (Jolliffe & Farrington, 2006) and the scale does not consistently relate to performance on emotion recognition tasks (e.g., Spreng, McKinnon, Mar, & Levine, 2009; Vachon & Lynam, 2016).

In contrast to the EC and PT scales, which are intended to map onto affective and cognitive empathy, respectively, the FN scale aims to measure tendencies to become imaginatively absorbed in the feelings and actions of characters in books and movies (Davis, 1983), and appears to capture a trait closely allied to Tellegen's absorption construct (Tellegen & Atkinson, 1974), with which it correlates substantially (Wickramasekera & Szlyk, 2003). The PD scale aims to measure "'self-oriented' feelings of personal anxiety and unease

in tense interpersonal settings" (Davis, 1983, p. 114). The PD scale correlates highly with trait negative emotionality (Hawk et al., 2013) and most of the items do not reference the presence of other people but refer generally to one's ability to function in "emergencies" and tense situations.

Researchers have adopted a variety of approaches to using the IRI and its scales in research. Because the FN and PD scale contents are not as typically associated with widespread conceptualizations of empathy, many studies analyze only the EC and PT scales (as consistent with the test designer's recommendations; Hatcher et al., 1994; Jolliffe & Farrington, 2006). Nevertheless, many researchers sum scores across all four subscales to create a "total score" (e.g., Decety, Lewis, & Cowell, 2015; Domes, Hollerbach, Vohs, Mokros, & Habermeyer, 2013), with some reporting only this total score and omitting data for the individual scales (e.g., Gill & Stickle, 2016; Wood, James, & Ciardha, 2014). Others have summed the EC and PD scales to create an aggregate measure of "affective empathy" and summed the PT and FN scales to create an aggregate measure of "cognitive empathy" (e.g., Bock & Hosser, 2014; Gabay, Shamay-Tsoory, & Goldfarb, 2016), a technique that has been criticized (Chrysikou & Thompson, 2016).

In view of the research we have reviewed, many published findings based on the aggregation of various IRI scales may be misleading. In this regard, the contribution of the PD scale is particularly suspect. Multiple studies have found that the PD scale correlates weakly with other empathy scales (e.g., Chrysikou & Thompson, 2016) and does not load on a higher order empathy factor (Chrysikou & Thompson, 2016; Hawk et al., 2013; Pulos, Elison, & Lennon, 2004). Although PD's low correlations with other empathy scales do not necessarily demonstrate that it is irrelevant to the construct, relying upon a total (or global) empathy score treats constituent scales as broadly equivalent indicators of the latent empathy construct and may conceal considerable differences in terms of the scales' relations with external criteria (e.g., Hengartner et al., 2014). Moreover, even though the PD scale tends not to be significantly associated with the EC and PT scales, Jordan, Amir, and Bloom (2016) found that it was substantially associated with measures of emotional and behavioral contagion, suggesting that it might reflect an isomorphic emotion-matching definition of empathy. Further investigation of the functioning of the PD scale is necessary to evaluate its construct validity as an index of empathy.

Although the 4-scale structure of the IRI has been almost exclusively used in past research, and has been validated in a number of studies (e.g., Hawk et al., 2013; Pulos et al., 2004), some investigations suggest that the items coalesce into a three-factor structure: a factor composed of the EC and PT items together, a factor composed of the FN items, and a factor composed of the PD items (e.g., Alterman, McDermott, Cacciola, & Rutherford, 2003; Siu & Shek, 2005). The lack of distinguishability of the EC and PT scales in past studies may further indicate that the PT scale should not be relied on as a measure of cognitive empathy; additional factorial investigation would be particularly valuable in this regard.

The ACME Scales

Vachon, Lynam, and Johnson (2014) observed meta-analytically that the IRI and other measures of empathy demonstrate very weak negative correlations with aggression, even though these two constructs have long been conceptualized as strongly related. They concluded that this finding suggests weak validity for the IRI and other empathy measures. Partly in response to these concerns, Vachon and Lynam (2016) developed the ACME, which contains three scales, as an attempt to fashion a measure that relates strongly to interpersonal malignancy. The Cognitive Empathy (CE) scale aims to measure the "ability to detect and understand emotional displays" (p. 136). The Affective Resonance (AR) scale aims to measure "emotional response in the observer that is congruent in valence to the target" (p. 136). The Affective Dissonance (AD) scale aims to measure "the experience of a contradictory emotional response—for example, taking pleasure in others' pain or feeling annoyed with others' happiness" (p. 136); it is reverse-scored so that higher scores indicate less affective dissonance.

The psychometric qualities of the ACME appear promising, especially its incremental validity above and beyond other empathy measures in statistically predicting a wide range of potentially maladaptive personality variables, such as aggression and externalizing behaviors (Vachon & Lynam, 2016). Vachon and Lynam did not, however, present their incremental validity results at the subscale level, which would provide helpful information regarding the validity of the different elements comprising the ACME.

The CE scale is composed primarily of items related to self-reported emotion recognition ability (e.g., "I have a hard time reading people's emotions"), although it includes a few items that may expand the construct by assessing respondents' inferences regarding the causes of others' behavior (e.g., "I can usually guess what's making someone angry"). Nevertheless, Vachon and Lynam (2016) reported that this scale did not correlate significantly with emotion recognition task performance, although the affective empathy scales of the ACME did. They similarly reported that neither the IRI PT scale nor the cognitive empathy scale of the BES (which also contains clear emotion recognition item content) were significantly predictive of empathic accuracy in these tasks. This lack of construct validity may not be a flaw in any particular item pool, but a general problem for any self-report questionnaire measure of emotion recognition abilities, given that, as multiple authors have pointed out, self-estimation of people's mind-reading

abilities does not appear to consistently track their actual abilities (e.g., Ickes, Stinson, Bissonnette, & Garcia, 1990; Realo et al., 2003). Though the evidence does not compellingly support the CE scale's validity as an indicator of individuals' cognitive empathic abilities, it may indirectly measure other aspects of empathic functioning.

The ACME's AR scale appears to represent a broader domain of affective empathy than the IRI EC scale, the latter of which relates predominantly to sympathy for others who are suffering. The AR scale includes a number of items similar to those of the IRI EC scale (e.g., "It makes me feel good to help someone in need"), but unlike the latter, it also includes items related to the *positive* feelings of others (e.g., "I enjoy making others happy"). In further contrast to the IRI, the AR scale also includes numerous items related to empathic restraint (e.g., "If I see that I am doing something that hurts someone, I will quickly stop"). Given that empathy is theorized to relate to both the inhibition of harming behaviors as well as the activation of caring behaviors, one might expect the AR scale to relate more strongly to interpersonal malignancy than the IRI EC scale. Nevertheless, Vachon and Lynam (2016) observed that the AR scale manifested nearly equivalent (negative) relations with aggression, psychopathy, and other externalizing variables as did the IRI EC scale. This surprising finding warrants further investigation.

The ACME AD scale constitutes a radical departure from prior indices of affective empathy. The items appear to relate to a cruel, resentful, misanthropic, antisocial disposition (e.g., "I get a kick out of making other people feel stupid"). Whereas empathy has generally been conceptualized on a dimension stretching from apathy to high empathy, this scale expands the dimension so that it runs from high maliciousness to high empathy, with apathy ostensibly falling in the middle of this expanded continuum (Vachon & Lynam, 2016). Perhaps unsurprisingly given its conceptual overlap with antagonism, Vachon and Lynam found that this scale bore significantly stronger negative correlations with aggression, psychopathy, and Machiavellianism than did the other empathy scales they examined. Furthermore, they found that the AD scale bore substantial negative relations with negative affect, emotion dysregulation, anxiety, depression, and anger, whereas other affective empathy scales did not.

As pointed out by Vachon and Lynam (2016), one potential objection to empathy measures is that they might be little more than measures of generalized negative emotionality or emotional lability. If empathy is defined as only emotion matching, such as a mother crying when her baby is crying, then it would probably be positively associated with these constructs. If empathy is defined as socially functional perspective-taking and caring, such as a mother calmly comforting her crying baby, then it might actually be associated with reduced negative emotionality (cf. Strathearn, Fonagy, Amico, & Montague, 2009), perhaps especially egocentricity-generating emotions of anxiety (cf. Todd, Forstmann, Burgmer, Brooks, & Galinsky, 2015). In either case, if the IRI or ACME scales demonstrate particularly strong correlations with negative emotionality, it could raise questions concerning their discriminant validity; the extant research raises such questions regarding the IRI PD scale and ACME AD scales.

The factorial validity of the ACME scales has not yet been extensively explored, but the initial investigations by Vachon and Lynam (2016) raise potential concerns regarding the role of reverse-worded (RW) items in the factor structure. RW items generally either employ a negating term (e.g., "I am not compassionate toward the feelings of others") or a conceptually opposite term (e.g., "I am cold-hearted toward the feelings of others"). Such items may in certain cases generate spurious factors ("artifactors") corresponding largely to the direction of item-keying.

The exploratory factor analysis (EFA) and CFA by Vachon and Lynam (2016) indicated, in two samples, that the factor structure of the ACME items was complicated by the presence of RW items, which comprise a majority of the items. The EFA they conducted indicated minor factors based on wording method, which they elected to treat as nonsubstantive. The CFA they conducted demonstrated inadequate fit for the three-factor structure, although model fit improved as a consequence of including two uncorrelated wording method factors. These wording method issues warrant further investigation.

Present Study

This article reports the first published validation study of the ACME scales in a sample not composed of undergraduate students (Vachon & Lynam, 2016) and also aims to provide a more comprehensive comparison of the ACME and IRI than in previous research. First, we used CFAs to test the replicability of the factor structure of the ACME scales presented by Vachon and Lynam (2016) as well as the factor structure of the IRI presented by Davis (1983), both of which we predicted would replicate in our sample. We complemented these confirmatory analyses with ESEM (Asparouhov & Muthén 2009; Marsh, Morin, Parker, & Kaur, 2014) to exploratorily examine the factor structure of both indices.

Second, we used multidimensional IRT models to examine the measurement precision of the ACME and IRI scales at varying trait levels. IRT is a psychometric modeling paradigm, encompassing a number of specific methods, which evaluates the response properties of individual items (e.g., difficulty, discrimination) in relationship to latent trait or ability dimensions (for a comprehensive review, see De Ayala, 2013). Classical test theory assumes that measurement precision is constant across varying trait levels, whereas IRT techniques test this assumption empirically. Past studies in other fields have used these

techniques to compare the measurement precision of competing questionnaires at varying trait levels, frequently observing poor measurement precision at low and/or high levels of various trait dimensions (e.g., Fraley, Waller, & Brennan, 2000; Olino et al., 2013). Given that clinicians and treatment outcome researchers may be interested in assessing substantial empathy deficits and changes following treatment (e.g., Michie & Lindsay, 2012; Palgi, Palgi, Ben-Ezra, & Shrira, 2014), and that researchers may be interested in measuring empathy within presumably lower trait level populations (e.g., prison settings; Young et al., 2015), the measurement precision of empathy scales at various trait levels should be assessed. To our knowledge, no studies have evaluated the constancy of measurement precision in empathy questionnaires; we made no a priori predictions regarding these analyses.

Third, a broad objective of our study was to compare the construct validity of the IRI and the ACME and their incremental validity above and beyond each other. We hypothesized that (a) both the ACME and IRI scales would demonstrate substantial construct validity by correlating negatively with interpersonally malignant personality traits, such as psychopathic meanness, sadism, and sexual objectification tendencies; but that (b) the ACME scales, even when excluding the AD scale, would demonstrate superior incremental validity in predicting a range of interpersonally malignant traits, above and beyond IRI scales; and (c) the IRI PD scale would demonstrate a pattern of correlations inconsistent with the other ACME and IRI scales, and would demonstrate higher correlations with measures of negative emotionality than with convergent and construct validity variables.

Method

Participants were 401 community participants (M_{age} = 35.5, SD_{age} = 11.0) recruited via Amazon Mechanical Turk (M-Turk); 46.5% were male and 53.0% were female (0.05% reported other than male or female). White participants comprised 71.8% the sample, whereas African Americans (7.2%), Hispanic or Latino/Latina (6.0%), Asian or Asian American (11.0%), and other race/ethnicities (3.9%) comprised far less of the sample. M-Turk samples have generally been observed to yield psychometrically high-quality data in psychological research, with more diverse population representation than undergraduate student samples (Miller, Crowe, Weiss, Maples-Keller, & Lynam, 2017).

Measures

Empathy

ACME (Vachon & Lynam, 2016) and *IRI* (Davis, 1983). All three ACME scales demonstrated high internal consistency, as ascertained by Cronbach's alpha and mean interitem correlation (MIC; ACME Cognitive Empathy, α = .89, MIC = .41; ACME Affective Resonance, α = .90, MIC = .42; ACME Affective Dissonance, α = .96, MIC = .69). The IRI scales demonstrated similarly high internal consistencies (IRI Empathic Concern, α = .88, MIC = .52; IRI Perspective-Taking, α = .86, MIC = .47; IRI Fantasy, α = .82, MIC = .41; IRI Personal Distress, α = .86, MIC = .47).

External Criteria

HEXACO Personality Inventory–60 (Ashton & Lee, 2009). The HEXACO–60 is a 60-item general personality questionnaire containing six factors, each composed of four lower order facets. It is similar to Big 5 personality measures, but also includes Honesty-Humility, which sometimes emerges in factor analyses of personality inventories of cross-cultural samples (Ashton & Lee, 2009; αs ranged in our sample from .77 [Emotionality] to .83 [Extraversion]).

The HEXACO Emotionality dimension comprises three lower order dimensions conceptually related to negative emotionality (Fearfulness, Anxiety, and Dependence) as well as a lower order Sentimentality dimension that assesses empathic sensitivity and meaningful emotional attachment to others. Because of the substantial content overlap between empathy and Sentimentality, we analyzed HEXACO Emotionality both with and without the Sentimentality content, hereafter referred to as HEXACO NSE (α = .72).

Multidimensional Personality Questionnaire–Short Form (MPQ-SF; Tellegen, 1982; Harkness, Tellegen, & Waller, 1995). The MPQ-SF is a 33-item self-report measure of general personality assessing 4 higher order dimensions (Positive Emotionality [PE], α = .87; Negative Emotionality [NE], α = .73; Constraint [CON], α = .57; and Absorption, α = .54) consisting of 11 lower order dimensions (PE: Wellbeing, Achievement, Social Potency, and Social Closeness; NE: Stress Reaction, Alienation, and Aggression; CON: Control, Harm Avoidance, and Traditionalism; Absorption is a standalone indicator).

Psychopathic Personality Inventory–Revised (PPI-R; Lilienfeld & Andrews, 1996; Lilienfeld & Widows, 2005). The PPI-R is a 154-item self-report measure of psychopathy intended largely for use with nonoffender samples. It contains 8 lower order subscales that generally coalesce into two higher order factors, Fearless Dominance (α = .92), consisting of Fearlessness, Social Influence, and Stress Immunity, and Self-centered Impulsivity (α = .96), consisting of Blame Externalization, Machiavellian Egocentricity, Carefree Nonplanfulness, and Rebellious Nonconformity. One subscale, Coldheartedness (α = .86), does not load highly onto these higher order factors and is a standalone index of callousness and lack of sentimentality.

Triarchic Psychopathy Measure (TriPM; Patrick, 2010). The TriPM is a 58-item measure of psychopathic traits that yields 3 factors: Boldness (α = .87), which is conceptually similar to PPI-R Fearless Dominance; Disinhibition (= .91), which is conceptually similar to PPI-R Self-Centered Impulsivity; and Meanness (α = .94), which is similar to PPI-R Coldheartedness but encompasses more content relevant to hostility and antagonism.

Inventory of Callous-Unemotional Traits (ICU; Frick, 2004). The ICU is a 24-item self-report measure of callous-unemotional traits, which are conceptually related to the affective deficits of psychopathy (Frick, 2004). It is composed of three factors (Essau, Sasagawa, & Frick, 2006): Callous (α = .89); Uncaring (α = .83); and Unemotional (α = .80).

Narcissistic Personality Inventory (NPI; Raskin & Terry, 1988). The NPI is the most widely used self-report measure of narcissistic traits, containing 40 items. It is composed of 3 factors (Ackerman et al., 2011): Leadership/Authority (α = .86); Grandiosity/Exhibitionism (α = .81); and Entitlement/Exploitativeness (α = .68).

Psychological Entitlement Scale (PES; Campbell, Bonacci, Shelton, Exline, & Bushman, 2004). The PES, composed of 8 items that yield a total score (α = .86), is intended to measure a "stable and pervasive sense that one deserves more and is entitled to more than others" (Campbell et al., 2004, p. 31), theoretically central to narcissism (e.g., Krizan & Herlache, 2017).

Personality Inventory for DSM-5–Brief Form (PID-5 BF; Krueger, Derringer, Markon, Watson, & Skodol, 2012). The PID-5 BF is a 25-item self-report measure of dimensions in the *Diagnostic and Statistical Manual of Mental Disorders*, 5th edition (American Psychiatric Association, 2013) alternative (Section 3) model of personality disorders. It yields scores for five dimensions of Negative Affect (α = .85), Detachment (α = .87), Antagonism (α = .89), Disinhibition (α = .89), and Psychoticism (α = .88). The Negative Affect dimension correlates positively with most personality disorder (PD) traits but particularly strongly ($r > .7$) with borderline PD and dependent PD traits (Thimm, Jordan, & Bach, 2016).

Varieties of Sadistic Tendencies Scale (VAST; Paulhus & Jones, 2014). The VAST is a 13 item self-report measure of sadistic tendencies, containing two subscales: Direct Sadism (α = .82, e.g., "I enjoy hurting people.") and Vicarious Sadism (α = .78, e.g., "In video games, I like the realistic blood spurts.").

Interpersonal Sexual Objectification Scale (ISOS; Kozee, Tylka, Augustus-Horvath, & Denchik, 2007). As a secondary index of behaviors and attitudes relevant to interpersonal callousness, we administered an amended version of the ISOS. The original version of the ISOS measures the extent to which individuals report experiencing sexual harassment and objectification. For this study, we amended the 21 items so that they assessed the extent to which the responding individual is the objectifier/harasser, rather than the target, of the objectification. Adapting the measure in this fashion afforded us an indicator of interpersonal malignancy in the sexual realm (α = .96).

Data Analysis

Missing Data. The prevalence of missing data were less than 2% for each item. Little's (1988) MCAR test was not statistically significant, which justified our performing single imputation using the expectation-maximization (EM) algorithm to impute missing data (Enders, 2001) for all composite scores used in external validity analyses.

Factor Analyses. To replicate the factor structure of the ACME established by Vachon and Lynam (2016), we employed factor analyses using the lavaan package (Rosseel, 2012) in R version 3.4, using the WLSMV estimator, which is most appropriate for ordinal data. Using this same method, Vachon and Lynam (2016) reported that the ACME three-scale structure demonstrated good fit, with three correlated substantive factors and two uncorrelated method factors (positive-coded and reverse-coded items, respectively). We similarly tested the model with (a) only the three trait factors and (b) with the three trait factors plus the two method factors, which were fixed to be uncorrelated with the substantive trait factors and with each other (this multitrait-multimethod factor analytic approach is a justified way of dealing with potential confounding of trait and wording/coding method covariance, e.g., Dimitrov, 2012). To investigate the factor structure of the IRI, we used a similar CFA approach. In addition to exploring the potential impact of RW items on the factor structure, we compared the four-factor structure with the alternative three-factor structure reported in a handful of prior studies (e.g., Siu & Shek, 2005).

In addition to the CFAs, we conducted ESEMs, a recently developed technique that combines features of EFA and CFA, as well as structural equation modeling (Asparouhov & Muthén, 2009; Marsh et al., 2014), in Mplus 7.0 (Muthén & Muthén, 1998-2012). These analyses, described further below, allowed us to examine the factor structure of the two measures in an exploratory manner.

Item Response Theory. To examine the measurement precision of the ACME and IRI scales, in their standard forms, at varying trait levels, we used graded response models (GRM; Samejima, 1969), which are commonly used for IRT purposes when dealing with ordinal polytomous response data (e.g., Likert-type scales). GRM is a two-parameter IRT

model (2PL), which examines both the polytomous form of item difficulty for each item, as well as each item's discrimination value (how reliably it discriminates between respondents higher versus lower on the trait dimension). GRM models generate item information curves, which estimate the measurement precision of an item depending on the level of trait or ability. The individual item information curves for items in a scale are summed to generate a test information function (TIF) curve. In the TIF curve, the conditional standard error of measurement for a given trait/ability value equals the inverse square root of the information level at that trait/ability value. TIFs can be equated with reliability; for instance, a TIF of 10 is equivalent to a marginal reliability of .9, whereas a TIF of 5 is equivalent to that of .8 (Embretson & Reise, 2000).

External Validity. In exploring the nomological networks associated with the ACME and IRI scales, we used the composite sum scores for the individual scales, given that most research employs these standard composite scores. We had .80 power to detect bivariate correlations at or above $r = .13$ at $p = .01$ and $r = .16$ at $p = .001$ (see Cohen, 1992). Given the inflated Type 1 error risk arising from the large number of tests, we report both levels of significance in the tables.

To compare the value of the ACME scales with those of the IRI scales, we performed hierarchical multiple regression analyses to ascertain the incremental validity of the ACME scales above and beyond the IRI scales, and vice versa, in statistically predicting variables in our data set theoretically characterized by interpersonal malignancy (TriPM Meanness, PPI-R Coldheartedness, PID-5 Antagonism, NPI Entitlement/Exploitativeness, VAST Direct, ICU Callous, and ISOS Total) given their relevance to empathy deficits. We entered either a single ACME scale or a combination of ACME scales (e.g., ACME CE and ACME AR) in the first block, and then an IRI counterpart, either a single scale or combination of scales, into the second block. We also conducted these analyses in reverse order, with an IRI scale or combination of scales in the first block, and ACME counterparts in the second block.

Results

Factor Structure of the ACME and IRI Scales

ACME. Using CFA, a correlated three-factor model, with AR, AD, and CE subscales, demonstrated inadequate fit (comparative fit index [CFI] = .90, Tucker-Lewis index [TLI] = .89, root mean square error of approximation [RMSEA] = .12, 95% confidence interval [CI; .12, .13]; $\chi^2 = 3898.76$, $df = 591$, $p < .001$). Replicating the results reported by Vachon and Lynam (2016), a five-factor CFA model with two method factors (positive and reverse wording) orthogonal to the three substantive factors (again, AR, AD, and CE) and to themselves demonstrated good fit (CFI = .98, TLI = .98, RMSEA = .05, 95% CI [.05, .06]; $\chi^2 = 1079.82$, $df = 555$, $p \leq .001$).

With regard to ESEMs of the ACME scales, Horn's (1965) parallel analysis indicated four dimensions underlying the item pool, but only three factors were associated with eigenvalues above 1 (the Kaiser criterion for factor extraction). The three-factor ESEM solution had an acceptable fit (CFI = .98, TLI = .98, RMSEA = .06, 95% CI [.06, .066], $\chi^2 = 1079.82$, $df = 555$, $p \leq .001$). Inspection of the estimated geomin-rotated factor matrix indicated substantial loadings of all AD items (all were RW) and RW AR and CE items (>.60) on the first factor. The second factor was mainly represented by both standard and RW CE items, with salient cross-loadings observed with regard to the latter category (>.5). The third factor was characterized by standard AR items, with cross-loadings of both standard CE items and RW AR items. All relevant items had salient loadings on their intended dimension only for the second factor, representing CE. The RW AR and CE items manifested significant cross-loadings on other factors. In sum, the three factors that emerged from the ESEM model did not clearly correspond to the intended dimensions of ACME. The four-factor ESEM solution had slightly improved fit indices (CFI = .99, TLI = .99, RMSEA = .05, 95% CI [.04, .05], $\chi^2 = 888.991$, $df = 492$, $p \leq .001$). The first two factors were similar to those in the three-factor ESEM. The third factor was again characterized by standard AR items, with weaker but salient (>.3) cross-loadings for RW AR items. The fourth factor had high cross-loadings from standard AR items.

In attempts to control for the method effects for both models, we respecified the models by freely estimated residual covariances between selected RW items; we used the modification indices to guide our model re-specifications. We obtained similar results. The cross-loadings with regard to the RW AR and CE items corroborated the CFA findings of Vachon and Lynam (2016) and our own CFA analyses, suggesting the likely existence of substantial wording method factors, either substantive or artifactual, in the factorial structure of the ACME items. A comparison of the CFA and ESEM models indicated that the ESEM results did not significantly improve in either fit or interpretability over the CFA model.

IRI. Turning to the IRI, the traditional four-factor structure (Davis, 1983) did not fit adequately in our sample (CFI = .87; TLI = .86; RMSEA = .11, 95% CI [.11, .12]; $\chi^2 = 1899.96$, $df = 344$, $p < .001$), nor did the alternative three-factor structure, with PT and EC collapsed into 1 factor (CFI = .85; TLI = .84; RMSEA = .12, 95% CI [.11, .12]; $\chi^2 = 2106.38.01$, $df = 347$, $p < .001$). Similar to our approach with the ACME, we tested a multitrait-multimethod CFA model with four correlated substantive factors (PT, EC, FN,

and PD) and two uncorrelated method (one positive wording and one negative wording). Including the potential method factors, this six-factor model demonstrated acceptable fit (CFI = .95; TLI = .94; RMSEA = .08, 95% CI [.07, .08]; χ^2 = 950.69, *df* = 316, $p < .001$); a multitrait multimethod CFA with 3-substantive factors and 2 method factors had marginal fit (CFI = .92; TLI = .90; RMSEA = .09, 95% CI [.09, .10]; χ^2 = 1309.25, *df* = 319, $p < .001$).

With regard to the ESEMs of the IRI items, Horn's parallel analysis indicated 5 dimensions, but only 4 were associated with an eigenvalue above 1. The four-factor ESEM solution had a marginal fit (CFI = .93, TLI = .91, RMSEA = .09, 95% CI [.08, .09], χ^2 = 1118.347, *df* = 272, $p \leq .001$). Inspection of the estimated geomin-rotated factor matrix indicated substantial loadings of all the EC and PT items on the first factor, both standard and RW. The second factor was represented by PD items, both standard and RW. The third factor was characterized by high factor and cross-loadings from RW EC, FN, PD, and PT items. The fourth factor was represented by the 7 FN items. The five-factor ESEM solution had slightly improved fit indices (CFI = .97, TLI = .96, RMSEA = .06, 95% CI [.06, .07], χ^2 = 652.99, *df* = 248, $p \leq .001$). The first factor had high loadings on the EC items; the second, PD; the third, all RW items; the fourth, PT, and finally, the fifth on the FS items. A comparison of the IRI CFA and ESEM models showed that the ESEM results did not significantly improve in fit or interpretability (Marsh et al., 2014).

Overall, our ESEM analyses pointed to the likelihood of substantial confounding of trait and method covariance for both the IRI and the ACME, much as Vachon and Lynam (2016) reported for the ACME. Similar to Vachon and Lynam's findings, though, the standard ACME and IRI structure models demonstrated acceptable fit once the two wording method factors were added to the CFA models. Following their lead, and consistent with multitrait-multimethod perspectives in scale development (e.g., Dimitrov, 2012), we used the typical factor structures of the two measures in our subsequent analyses, employing composite scores. This approach allowed us to ascertain the item properties and external correlates of these two measures as they have typically been used in prior studies and in clinical practice.

Measurement Precision at Varying Trait Levels

Supplemental Figures 1 to 7, available with the online version of the article, display the test information curves of the seven scales, based on the CFA models estimated for ACME and IRI. Other than the CE and the PD scales to some extent, all the empathy scales displayed low precision at higher empathy trait levels over 0.5 to 1.5. Their accuracy sharply decreases as the trait level increases and the standard error of measurement rises from an average of 0.25 in the low trait level to an average of 0.7 for the trait level above 2.0. In sum, the item pools for most of the scales of both measures do not appear to measure empathy with adequate precision with regard to individuals with markedly elevated empathic traits.

Relations Between ACME and IRI Scales

We next turn to the correlations between the ACME and IRI scales, which can be found in Table 1. All three ACME scales, the IRI EC scale, and the IRI PT scales intercorrelated positively. The IRI FN scale was not significantly correlated with the ACME AD scale but was positively correlated with all other scales. The IRI PD scale was weakly negatively correlated with most of the other empathy scales, weakly positively with the IRI FN scale, and negligibly with the IRI EC scale (ironically, the scale with which it is frequently aggregated to create an "affective empathy" composite, e.g., Gabay et al., 2016). Although the IRI PT scale is often treated as a "cognitive" empathy variable, it demonstrated larger positive correlations with the major affective empathy scales, namely, IRI EC and ACME AR, than with the ACME CE scale (smallest Steiger's (1980) z = 3.65, $p < .001$), pointing to problems with its discriminant validity.

Relations Between Empathy and External Criteria

Relations With Broad-Band Personality Dimensions. Correlations between the ACME (and the IRI) and both the HEXACO and MPQ dimensions are presented in Table 2. The ACME scales and the IRI EC and PT scales were positively correlated with all six HEXACO dimensions. The pattern of correlations was similar for the ACME scales and the IRI EC and PT scales, which can be interpreted as additional evidence of the ACME's convergent validity. The FN scale demonstrated similar correlations for most of the HEXACO dimensions but was not significantly correlated with Honesty/Humility or Agreeableness, indicating a lack of convergent validity with the other empathy scales. The IRI PD scale was more highly correlated with HEXACO Emotionality (r = .62) and HEXACO NSE (r = .59), than it was with the other HEXACO dimensions (all Steiger's z tests, $p < .001$). The IRI PD scale was negatively correlated with Honesty-Humility, Extraversion, Agreeableness, Conscientiousness, and Openness.

All the affective empathy scales, as well as IRI PT, were significantly and positively correlated with MPQ Constraint, but ACME CE, IRI FN, and IRI PD were not. Absorption demonstrated small correlations with all empathy scales, except for ACME AR, with which it was not significantly correlated, and IRI FN, with which it demonstrated a medium positive correlation. As expected, absorption was

Table 1. Correlations Between ACME and IRI Scales.

	ACME CE	ACME AR	ACME AD	IRI PT	IRI EC	IRI FN	IRI PD
ACME CE	—						
ACME AR	**.47**	—					
ACME AD	**.29**	**.74**	—				
IRI PT	**.43**	**.58**	**.39**	—			
IRI EC	**.44**	**.77**	**.52**	**.66**	—		
IRI FN	**.33**	**.31**	.09	**.35**	**.42**	—	
IRI PD	−.12	*−.14*	**−.24**	*−.15*	−.08	.13	—

Note. $N = 401$. Bolded is $p < .001$, italicized is $p < .01$. ACME = Affective and Cognitive Measure of Empathy; CE = Cognitive Empathy; AR = Affective Resonance; AD = Affective Dissonance; IRI = Interpersonal Reactivity Index; PT = Perspective Taking; EC = Empathic Concern; FN = Fantasy, PD = Personal Distress.

Table 2. Correlations Between Empathy Measures and General Personality Indices.

	HEXACO							MPQ			
	H	E	X	A	C	O	NSE	PE	NE	CON	ABS
ACME CE	**.21**	*.15*	**.25**	**.22**	**.35**	**.36**	.03	**.20**	**−.21**	.07	**.18**
ACME AR	**.49**	**.26**	**.23**	**.45**	**.51**	**.45**	.10	.08	**−.45**	**.24**	.06
ACME AD	**.49**	.07	.10	**.36**	**.56**	**.35**	−.05	**−.21**	**−.56**	**.25**	**−.20**
IRI PT	**.41**	*.16*	**.30**	**.62**	**.41**	**.41**	.02	**.25**	**−.40**	**.23**	**.18**
IRI EC	**.51**	**.36**	**.29**	**.46**	**.40**	**.41**	*.16*	**.22**	**−.36**	**.24**	*.13*
IRI FN	.01	**.33**	.12	.08	.13	**.41**	**.25**	**.20**	.01	.02	**.41**
IRI PD	−.12	**.61**	**−.42**	**−.23**	**−.27**	*−.17*	**.59**	**−.17**	**.49**	.05	*.13*

Note. $N = 401$. Bolded indicates $p < .001$, italicized indicates $p < .01$. ACME = Affective and Cognitive Measure of Empathy; CE = Cognitive Empathy, AR = Affective Resonance, AD = Affective Dissonance; IRI = Interpersonal Reactivity Index; PT = Perspective Taking; EC = Empathic Concern; FN = Fantasy; PD = Personal Distress; HEXACO = HEXACO Personality Inventory; H = Honesty/Humility; E = Emotionality; X = Extraversion; A = Agreeableness; C = Conscientiousness; O = Openness; NSE = Nonsentimental Emotionality (i.e., a modification of HEXACO E such that items from the Sentimentality subscale are elided); MPQ = Multidimensional Personality Questionnaire; PE = Positive Emotionality; NE = Negative Emotionality; CON = Constraint; ABS = Absorption.

more strongly correlated with IRI FN than with any other empathy scale (smallest Steiger's $z = 3.33$, $p < .001$).

Relations With Personality Disorders Features or Interpersonal Malignancy. All correlations with the range of convergent and discriminant personality correlates are presented in Table 3. The scales most clearly related to affective empathy (ACME AR, ACME AD, and IRI EC) all correlated negatively and robustly with all interpersonal malignancy variables, psychological entitlement and also all five personality dimensions in the PID-5. These primary affective empathy scales were not robustly associated with psychopathy or narcissism variables primarily measuring confidence/social potency (PPI-R FD, TriPM Boldness, and NPI Leadership/Authority). Though ACME AD generally demonstrated the highest correlations with interpersonal malignancy and PD features, the ACME AR scale was also typically more strongly correlated with these variables than the IRI EC scale. The Steiger's z comparing ACME AR to the next strongest correlated IRI scale was significant, at $p < .001$, for: TriPM Meanness, TriPM Disinhibition, PPI-R SCI, ICU Callous, PES Entitlement, VAST Direct Sadism, ISOS Total, PID-5 Antagonism, PID-5 Disinhibition, PID-5 Detachment, PID-5 Psychoticism, and PID-5 Negative Affect. In contrast, the IRI EC scale only demonstrated a significantly stronger relationship than the ACME AR scale with PPI-R CH and ICU Unemotional.

The pseudo-cognitive scales (IRI PT and ACME CE) displayed similar nomological networks as did the affective empathy scales, although the correlational associations were generally not as pronounced as those demonstrated by the ACME AR and AD scales. This pattern of convergent validity partially supports their relationships to the empathy construct, in some fashion, even if neither should be used as a proxy for actual cognitive empathy ability.

The IRI FN scale was, in general, only weakly negatively associated with interpersonal malignancy and personality disorder feature variables. The IRI PD scale was generally positively associated with interpersonal malignancy and personality disorder features, but it was robustly negatively correlated with the confidence/social potency variables.

Table 3. Correlations Between Empathy Measures and Indices of Personality Disorder Features and Interpersonal Malignancy.

	ACME			IRI			
	CE	AR	AD	PT	EC	FN	PD
TriPM							
Boldness	**.21**	.03	.01	*.15*	.07	.00	**−.62**
Meanness	**−.35**	**−.80**	**−.82**	**−.56**	**−.71**	**−.20**	**.19**
Disinhibition	**−.28**	**−.61**	**−.77**	**−.42**	**−.49**	−.07	**.35**
ICU							
Callous	**−.37**	**−.72**	**−.83**	**−.40**	**−.57**	**−.19**	**.20**
Uncaring	**−.45**	**−.61**	**−.36**	**−.56**	**−.59**	**−.25**	**.19**
Unemotional	**−.36**	**−.28**	−.13	**−.28**	**−.38**	**−.22**	.12
PPI-R							
FD	.06	−.11	**−.19**	.04	−.07	−.06	**−.54**
SCI	**−.27**	**−.63**	**−.75**	**−.46**	**−.53**	−.06	**.29**
C	**−.35**	**−.63**	**−.28**	**−.57**	**−.72**	**−.38**	**−.19**
NPI							
L/A	.10	*−.17*	**−.29**	−.06	*−.13*	−.03	**−.27**
GE	−.02	**−.25**	**−.38**	*−.15*	*−.18*	.04	.01
E/E	*−.16*	**−.48**	**−.58**	**−.29**	**−.41**	−.09	.09
VAST							
Direct	**−.27**	**−.67**	**−.83**	**−.40**	**−.49**	−.07	**.20**
Vicarious	**−.18**	**−.52**	**−.64**	**−.33**	**−.47**	−.13	.00
PID-5							
Antagonism	**−.27**	**−.67**	**−.81**	**−.39**	**−.49**	−.11	**.23**
Disinhibition	**−.29**	**−.56**	**−.75**	**−.35**	**−.40**	−.10	**.28**
Detachment	**−.35**	**−.57**	**−.57**	**−.39**	**−.49**	**−.20**	**.36**
Psychoticism	**−.24**	**−.52**	**−.71**	**−.29**	**−.39**	.02	**.33**
Negative Affect	−.13	**−.30**	**−.47**	**−.23**	**−.18**	.11	**.59**
ISOS	**−.22**	**−.61**	**−.80**	**−.31**	**−.44**	−07	**.18**
PES	**−.28**	**−.61**	**−.66**	**−.41**	**−.52**	−.13	**.18**

Note. Bolded is $p < .001$, italicized is $p < .01$. ACME = Affective and Cognitive Measure of Empathy; CE = Cognitive Empathy; AR = Affective Resonance; AD = Affective Dissonance;IRI = Interpersonal Reactivity Index; PT = Perspective Taking; EC = Empathic Concern; FN = Fantasy; PD = Personal Distress; TriPM = Triarchic Psychopathy Measure; ICU = Inventory of Callous-unemotional Traits; PPI-R = Psychopathic Personality Inventory–Revised; FD = Fearless Dominance; SCI = Self-centered Impulsivity; C = Coldheartedness; NPI = Narcissistic Personality Inventory; L/A = Leadership/Authority; GE = Grandiose Exhibitionism; E/E = Entitlement/Exploitativeness; VAST = Varieties of Sadistic Tendencies scale; Direct = Direct Sadism; Vicarious = Vicarious Sadism; PID-5 = Personality Inventory for *DSM-5*–Brief Form; ISOS = Interpersonal Sexual Objectification Scale; PES = Psychological Entitlement Scale.

Discriminant Validity From Broad-Band Emotional Dimensions. As shown in Table 2, only ACME AD and IRI PD were negatively correlated with MPQ Positive Emotionality; all other scales, except for ACME AR, were positively correlated with it. All empathy scales were negatively correlated with MPQ Negative Emotionality, except for IRI FN, which was not significantly correlated with it, and IRI PD, which was substantially positively correlated with it. Most empathy scales were not significantly related to HEXACO NSE, but IRI EC and FN both exhibited small positive correlations with it, and IRI PD demonstrated a large positive correlation with it. Most empathy scales demonstrated small negative correlations with PID-5 Negative Affect (see Table 3), but ACME AD demonstrated a medium negative correlation with it, IRI FN demonstrated a small positive correlation with it, and IRI PD demonstrated a large positive correlation with it.

Our results suggest that the IRI PD scale is more a measure of negative emotionality than of empathy. The ACME AD scale correlated more strongly with PID-5 Negative Affect (smallest Steiger's $z = 5.54$, $p < .001$) and MPQ NE (smallest Steiger's $z = 4.00$, $p < .001$) than any other empathy scale aside from IRI PD.

Incremental Validity of the ACME Above and Beyond the IRI, and Vice Versa

For full incremental validity results, see Supplemental Table 1, available with the online version of the article. The full ACME scales (CE + AR + AD) demonstrated a high level of incremental validity above and beyond the core IRI scales (EC + PT) in statistically predicting our seven selected interpersonal malignancy variables (average

ΔR^2 = .30). More tellingly, although weaker than the statistical effects of the full ACME scales, the combination of only the ACME CE and AR scales also demonstrated medium incremental validity above and beyond the IRI EC and PT scales in statistically predicting interpersonal malignancy (average ΔR^2 = .13).

By comparison, when the core IRI scales were entered in the second block, they yielded minimal incremental validity above the combination of ACME CE and AR in statistically predicting interpersonal malignancy (average ΔR^2 = .03). Of the seven outcome variables, they only offered meaningful incremental validity in predicting PPI-R CH (ΔR^2 = .14)

The ACME AR scale by itself demonstrated medium incremental validity above and beyond IRI EC in statistically predicting interpersonal malignancy (average ΔR^2 = .13). By comparison, the IRI EC scale produced weak incremental validity in statistically predicting interpersonal malignancy (average ΔR^2 = .03), only demonstrating meaningful incremental validity in statistically predicting PPI-R CH (ΔR^2 = .14).

Discussion

Our analyses indicate that both the IRI and the ACME scales have some strengths, but also substantial limitations. Our results: leave some lingering concerns regarding the factorial validity of both questionnaires; indicate that both lack measurement precision at higher trait levels; and raise concerns regarding the construct validity of particular scales. Overall, though, our analyses indicate that the ACME affords substantial advantages over the IRI in relating to interpersonally malignant traits, consistent with Vachon and Lynam's (2016) goals in crafting it.

Factor Structures of the ACME and IRI

Consistent with the results obtained by Vachon and Lynam, our CFAs and ESEMs indicated that the factor structure of the ACME scales is significantly complicated by the mixture of RW and standard items. Our results revealed similar problems for the IRI, though it has a lower proportion of RW items. Many scales employ RW items to combat acquiescence bias. At the same time, though, this practice may introduce wording method confounds. Factor analyses of scales with some RW items frequently indicate the presence of method covariance obscuring or confounding substantive covariance (e.g., Brown, 2003; Roszkowski & Soven, 2010). The factor analysis distortion caused by RW items appears to be relatively sensitive to differences in responding styles between study samples. For instance, even a small amount of careless responding in a sample can cause wording method factors to manifest in factor analyses (e.g., Woods, 2006). CFAs of scales with RW items often fail to provide adequate fit unless the model is adjusted to account for wording method variance (e.g., Woods, 2006).

Nevertheless, discrepancies between RW and standardly worded items may sometimes reflect genuine personality variance (rather than pure method variance), especially when the items relate to social desirability or self-esteem (DiStefano & Motl, 2009). For instance, query items such as "I am an empathic person" and "I am not an empathic person" may not be direct opposites of each other, as defensiveness or fear of negative evaluation may be more related to the latter. In other words, agreeing to a socially negative statement about oneself may not be the precise methodological opposite of disagreeing about a positive statement about oneself.

Our analyses do not allow us to disentangle whether the wording method factors that emerge with the ACME and IRI items reflect a method artifact, a substantive (personality) dimension, or a mixture of both. Nevertheless, in concert with the findings of Vachon and Lynam (2016), our findings encourage future research on wording method effects, especially given that they can complicate other analyses. After accounting for these wording method effects in our CFAs, the three-scale structure of the ACME presented by Vachon and Lynam and the four-scale structure of the IRI presented by Davis (1983) were replicated in our sample.

Measurement Precision of the ACME and IRI Scales

In our IRT analyses, a strong general pattern held: The ACME and IRI scales appear to be effective in detecting moderate and low levels of empathic traits, but lack measurement precision at high levels of these traits. Because this is the first study to examine the IRT properties of these empathy measures, future replication is needed. If further research comes to similar conclusions, this would indicate that these measures may be reliable when evaluating empathy deficits and their change over time, as well as in research with lower empathy subject samples. If future research corroborates our finding that these measures lack reliability at higher trait levels, however, the use of these measures in high-empathy populations, such as individuals in helping professions, may be methodologically problematic.

Construct and Incremental Validity

Although the IRI's main affective empathy scale, the EC scale, appears to possess a highly similar nomological network to that of the ACME AR scale, the AR scale demonstrated substantial incremental validity benefits in comparison, at least in terms of traits related to meanness, antagonism, and other aspects of interpersonal malignancy. We observed similar results when examining the incremental validity of combinations of the ACME scales, above and beyond combinations of the IRI scales. These findings

suggest that Vachon and Lynam (2016) successfully devised a measure that more effectively correlates with malignant empathy deficits than does the IRI. The IRI, however, performed better than the ACME in statistically predicting PPI-R Coldheartedness, indicating that it may still offer value not captured by the ACME scales (but see other promising data in Vachon & Lynam, 2016, indicating heightened aversion to emotionally negative stimuli as a more pronounced correlate of the ACME AR scale than of the IRI EC scale). Because the Coldheartedness scale is more a measure of passive emotional detachment than of active antagonism, this result raises the possibility that the IRI is a better marker of this feature than is the ACME.

Our results also indicate that the ACME AD scale appears to more strongly measure interpersonal malignancy and emotional hostility than it does general empathy. The AD scale correlated more robustly with interpersonal malignancy variables than with the other empathy scales, reflecting the AD scale's strong saturation with reversed antagonism and reversed conscientiousness. Furthermore, the AD scale correlated more strongly with both MPQ Negative Emotionality and PID-5 Negative Affect than did any of the other empathy scales aside from the IRI PD scale. This disproportionate tilt toward negative emotionality conceptually replicates findings by Vachon and Lynam (2016) regarding the AD scale.

Our results also offer compelling evidence that the IRI PD scale is much more a measure of negative emotionality than of empathy. This scale is (a) much more strongly correlated with negative emotionality than with the other empathy scales; (b) correlated with general personality traits in opposing or otherwise divergent ways compared with other empathy scales; and (c) in general correlated *positively* with measures characterized by interpersonal malignancy, the opposite of what one would expect from an empathy scale. As a consequence, many published analyses that have included this scale in either a "total" IRI score (e.g., Flight & Forth, 2007; Glass, Moody, Grafman, & Krueger, 2016) or aggregated it with the IRI Empathic Concern scale to create an "affective empathy" variable (e.g., Dziobek et al., 2011; Shamay-Tsoory et al., 2009) may be seriously misleading. Our results, in conjunction with similarly critical analyses (e.g., Chrysikou & Thompson, 2016), argue against the widespread practice of aggregating this scale with other scales to create "total" or composite empathy variables.

Although not as starkly troubling as for the PD scale, our results corroborate the view that the IRI FN scale possesses only limited construct validity as an empathy measure. Although it demonstrated some convergent validity with other empathy scales, our data suggest that it may be more a measure of trait absorption than of empathy. Given the low internal consistency of our short-form MPQ Absorption scale, though, these results should be replicated using the full scale.

Our findings also elucidate the role of negative emotionality in empathy as operationalized by the ACME and IRI. The main empathy scales (ACME AR, ACME CE, IRI EC, and IRI PT) were not robustly associated with the HEXACO NSE variable, but they were robustly negatively correlated with MPQ NE and PID-5 Negative Affect. This finding could indicate that empathy is negatively associated with feeling disconnected from others, stress-related irritability, and other emotional correlates of interpersonal friction but not substantially associated with less interpersonally charged aspects of negative emotionality, such as worry-proneness. Future research, using more fine-grained measures of specific aspects of negative emotionality, may further illuminate this aspect of the empathy construct.

Although further research is necessary to conceptually replicate our findings, we tentatively propose that researchers should prefer the ACME AR over the IRI EC scale when measuring affective empathy but should administer both if possible, pending further comparative validation research. Although the ACME CE scale is, at least facially, more related to cognitive empathy ability than is the IRI PT scale, it remains unclear how effectively it measures such abilities. We advise that researchers not use the ACME CE nor the IRI PT as proxies for cognitive empathy unless adequate convergent validity with behavioral results emerges. The ACME AD scale should probably be used with caution as a measure of empathy, although it may detect some of the more interpersonally malignant correlates of this construct.

Limitations and Future Directions

One limitation of this study is that we did not administer the EQ (Baron-Cohen & Wheelwright, 2004). This self-report measure of empathy, which can be parsed into distinguishable cognitive and affective empathy scales (Lawrence, Shaw, Baker, Baron-Cohen, & David, 2004), is in much the same theoretical vein as the IRI and the ACME. Further research is needed to investigate the value of the ACME relative to the EQ.

Another limitation is that we did not administer task assessments of emotion recognition or other features of cognitive empathy. As highlighted by Vachon and Lynam (2016), self-report measures of cognitive empathy abilities may reflect people's appraisal of their own accuracy rather than such accuracy per se. Unless research establishes that individuals are valid judges of their own cognitive empathy abilities, the ACME CE scale and other comparable scales should not be relied on to quantify this aspect of empathy.

In sum, the ACME represents a substantial step forward in empathy research, as it has attempted to broaden the content domain of empathy measurement. Such attempts can, however, be expanded even further. For instance, inspired by the IRI FN scale, future research should investigate the measurement of potential empathic qualities involved in appreciating

the stories of others, or of listening skills and tendencies. Although the FN scale focuses on absorption in books and movies, human empathy and human stories have existed far longer than the printing press and Hollywood, and person-to-person communications may be a more fruitful domain for empathy research and measurement.

Similarly, as suggested by the PD scale, empathy can be difficult and even distressing, and susceptibilities to empathic distress or empathic fatigue may be important aspects of empathic functioning. Whereas the PD scale references "emergencies" and ambiguous tense situations, future research could investigate this process more explicitly in interpersonal empathic encounters, such as listening to a friend talk about the death of a loved one or an acquaintance open up about a recent traumatic experience. These are merely a few of the potential domains into which multidimensional empathy measures could be extended. Before the contours of the empathy construct can be more comprehensively mapped and appropriately measured, additional effort should be invested in expanding the measurement content domain to better understand the boundaries of this still poorly understand construct (cf. Clark & Watson, 1995; Loevinger, 1957).

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

Supplementary material for this article is available online.

References

Ackerman, R. A., Witt, E. A., Donnellan, M. B., Trzesniewski, K. H., Robins, R. W., & Kashy, D. A. (2011). What does the Narcissistic Personality Inventory really measure? *Assessment*, *18*, 67-87.

Alterman, A. I., McDermott, P. A., Cacciola, J. S., & Rutherford, M. J. (2003). Latent structure of the Davis Interpersonal Reactivity Index in methadone maintenance patients. *Journal of Psychopathology and Behavioral Assessment*, *25*, 257-265.

American Psychiatric Association. (2013). *Online assessment measures: The Personality Inventory for DSM-5–Brief Form (PID-5-BF)–Adult*. Retrieved from https://www.psychiatry.org/psychiatrists/practice/dsm/educational-resources/assessment-measures

Ashton, M. C., & Lee, K. (2009). The HEXACO-60: A short measure of the major dimensions of personality. *Journal of Personality Assessment*, *91*, 340-345.

Asparouhov, T., & Muthén, B. O. (2009). Exploratory structural equation modeling. *Structural Equation Modeling*, *16*, 397-438.

Baron-Cohen, S. (2016, December 30). Empathy is good, right? A new book says we're better off without it. *New York Times*. Retrieved from https://www.nytimes.com/2016/12/30/books/review/against-empathy-paul-bloom.html

Baron-Cohen, S., & Wheelwright, S. (2004). The empathy quotient: An investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders*, *34*, 163-175.

Bird, G., & Viding, E. (2014). The self to other model of empathy: Providing a new framework for understanding empathy impairments in psychopathy, autism, and alexithymia. *Neuroscience & Biobehavioral Reviews*, *47*, 520-532.

Bloom, P. (2016). *Against empathy: The case for rational compassion*. New York, NY: Ecco.

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

Bock, E. M., & Hosser, D. (2014). Empathy as a predictor of recidivism among young adult offenders. *Psychology, Crime & Law*, *20*, 101-115.

Brown, T. A. (2003). Confirmatory factor analysis of the Penn State Worry Questionnaire: Multiple factors or method effects? *Behaviour Research and Therapy*, *41*, 1411-1426.

Campbell, W. K., Bonacci, A. M., Shelton, J., Exline, J. J., & Bushman, B. J. (2004). Psychological entitlement: Interpersonal consequences and validation of a self-report measure. *Journal of Personality Assessment*, *83*, 29-45.

Chrysikou, E. G., & Thompson, W. J. (2016). Assessing cognitive and affective empathy through the Interpersonal Reactivity Index: An argument against a two-factor model. *Assessment*, *23*, 769-777.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*, 309-319.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159.

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

De Ayala, R. J. (2013). The IRT tradition and its applications. *Oxford Handbook of Quantitative Methods: Foundations*, *1*, 144-169.

Decety, J., Lewis, K. L., & Cowell, J. M. (2015). Specific electrophysiological components disentangle affective sharing and empathic concern in psychopathy. *Journal of Neurophysiology*, *114*, 493-504.

Decety, J., & Michalska, K. J. (2010). Neurodevelopmental changes in the circuits underlying empathy and sympathy from childhood to adulthood. *Developmental Science*, *13*, 886-899.

Dimitrov, D. M. (2012). *Statistical methods for validation of assessment scale data in counseling and related fields*. Alexandria, VA: American Counseling Association.

DiStefano, C., & Motl, R. W. (2009). Personality correlates of method effects due to negatively worded items on the Rosenberg Self-Esteem scale. *Personality and Individual Differences*, *46*, 309-313.

Domes, G., Hollerbach, P., Vohs, K., Mokros, A., & Habermeyer, E. (2013). Emotional empathy and psychopathy in offenders: An experimental study. *Journal of Personality Disorders*, *27*, 67-84.

Dziobek, I., Preißler, S., Grozdanovic, Z., Heuser, I., Heekeren, H. R., & Roepke, S. (2011). Neuronal correlates of altered

empathy and social cognition in borderline personality disorder. *Neuroimage*, *57*, 539-548.

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

Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling*, *8*, 128-141.

Essau, C. A., Sasagawa, S., & Frick, P. J. (2006). Callous-unemotional traits in a community sample of adolescents. *Assessment*, *13*, 454-469.

Flight, J. I., & Forth, A. E. (2007). Instrumentally violent youths: The roles of psychopathic traits, empathy, and attachment. *Criminal Justice and Behavior*, *34*, 739-751.

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology*, *78*, 350-365.

Frick, P. J. (2004). *The Inventory of Callous–Unemotional Traits* (Unpublished rating scale). University of New Orleans, New Orleans, LA.

Gabay, Y., Shamay-Tsoory, S. G., & Goldfarb, L. (2016). Cognitive and emotional empathy in typical and impaired readers and its relationship to reading competence. *Journal of Clinical and Experimental Neuropsychology*, *38*, 1131-1143.

Gill, A. D., & Stickle, T. R. (2016). Affective differences between psychopathy variants and genders in adjudicated youth. *Journal of Abnormal Child Psychology*, *44*, 295-307.

Glass, L., Moody, L., Grafman, J., & Krueger, F. (2016). Neural signatures of third-party punishment: Evidence from penetrating traumatic brain injury. *Social Cognitive and Affective Neuroscience*, *11*, 253-262.

Harkness, A. R., Tellegen, A., & Waller, N. (1995). Differential convergence of self-report and informant data for Multidimensional Personality Questionnaire traits: Implications for the construct of negative emotionality. *Journal of Personality Assessment*, *64*, 185-204.

Hatcher, S. L., Nadeau, M. S., Walsh, L. K., Reynolds, M., Galea, J., & Marz, K. (1994). The teaching of empathy for high school and college students: Testing Rogerian methods with the Interpersonal Reactivity Index. *Adolescence*, *29*, 961-975.

Hawk, S. T., Keijsers, L., Branje, S. J., Graaff, J. V. D., Wied, M. D., & Meeus, W. (2013). Examining the Interpersonal Reactivity Index (IRI) among early and late adolescents and their mothers. *Journal of Personality Assessment*, *95*, 96-106.

Hengartner, M. P., De Fruyt, F., Rodgers, S., Mueller, M., Roessler, W., & Ajdacic-Gross, V. (2014). An integrative examination of general personality dysfunction in a large community sample. *Personality and Mental Health*, *8*, 276-289.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179-185.

Ickes, W., Stinson, L., Bissonnette, V., & Garcia, S. (1990). Naturalistic social cognition: Empathic accuracy in mixed-sex dyads. *Journal of Personality and Social Psychology*, *59*(4), 730-742.

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

Jordan, M. R., Amir, D., & Bloom, P. (2016). Are empathy and concern psychologically distinct? *Emotion*, *16*, 1107-1116.

Keltner, D., & Haidt, J. (1999). Social functions of emotions at four levels of analysis. *Cognition & Emotion*, *13*, 505-521.

Keltner, D., & Kring, A. M. (1998). Emotion, social function, and psychopathology. *Review of General Psychology*, *2*, 320-342.

Kozee, H. B., Tylka, T. L., Augustus-Horvath, C. L., & Denchik, A. (2007). Development and psychometric evaluation of the interpersonal sexual objectification scale. *Psychology of Women Quarterly*, *31*, 176-189.

Krizan, Z., & Herlache, A. D. (2017). The narcissism spectrum model: A synthetic view of narcissistic personality. *Personality and Social Psychology Review*, *22*(1), 3-31.

Lawrence, E. J., Shaw, P., Baker, D., Baron-Cohen, S., & David, A. S. (2004). Measuring empathy: Reliability and validity of the Empathy Quotient. *Psychological Medicine*, *34*, 911-920.

Lilienfeld, S. O., & Widows, M. (2005). *Professional manual for the Psychopathic Personality Inventory–Revised (PPI-R)*. Lutz, FL: Psychological Assessment Resources.

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*, 1198-1202.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*, 635-694.

Marsh, H. W., Morin, A. J. S., Parker, P., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology*, *10*, 85-110.

Michie, A. M., & Lindsay, W. R. (2012). A treatment component designed to enhance empathy in sex offenders with an intellectual disability. *British Journal of Forensic Practice*, *14*(1), 40-48.

Miller, J. D., Crowe, M., Weiss, B., Maples-Keller, J. L., & Lynam, D. R. (2017). Using online, crowdsourcing platforms for data collection in personality disorder research: The example of Amazon's Mechanical Turk. *Personality Disorders: Theory, Research, and Treatment*, *8*, 26-34.

Muthén, L. K., & Muthén, B. O. (1998-2012). *MPlus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Olino, T. M., Yu, L., McMakin, D. L., Forbes, E. E., Seeley, J. R., Lewinsohn, P. M., & Pilkonis, P. A. (2013). Comparisons across depression assessment instruments in adolescence and young adulthood: An item response theory study using two linking methods. *Journal of Abnormal Child Psychology*, *41*, 1267-1277.

Palgi, S., Palgi, Y., Ben-Ezra, M., & Shrira, A. (2014). "I will fear no evil, for I am with me": Mentalization-oriented intervention with PTSD patients. A case study. *Journal of Contemporary Psychotherapy*, *44*, 173-182.

Patrick, C. J. (2010). *Triarchic Psychopathy Measure (TriPM)*. Retrieved from https://www.phenxtoolkit.org/index.php?pageLink=browse.protocoldetails&id=121601

Paulhus, D. L., & Jones, D. N. (2014). Measures of dark personalities. In G. J. Boyle, D. H. Saklofske & G. Matthews (Eds.), *Measures of personality and social psychological constructs* (pp. 562-594). San Diego, CA: Academic Press.

Pulos, S., Elison, J., & Lennon, R. (2004). The hierarchical structure of the Interpersonal Reactivity Index. *Social Behavior and Personality: An International Journal*, *32*, 355-359.

Raskin, R., & Terry, H. (1988). A principal-components analysis of the Narcissistic Personality Inventory and further evidence

of its construct validity. *Journal of Personality and Social Psychology*, *54*, 890-902.

Realo, A., Allik, J., Nõlvak, A., Valk, R., Ruus, T., Schmidt, M., & Eilola, T. (2003). Mind-reading ability: Beliefs and performance. *Journal of Research in Personality*, *37*, 420-445.

Reniers, R. L., Corcoran, R., Drake, R., Shryane, N. M., & Völlm, B. A. (2011). The QCAE: A questionnaire of cognitive and affective empathy. *Journal of Personality Assessment*, *93*, 84-95.

Rogers, C. R. (1958). The characteristics of a helping relationship. *Journal of Counseling & Development*, *37*, 6-16.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1-36.

Roszkowski, M. J., & Soven, M. (2010). Shifting gears: Consequences of including two negatively worded items in the middle of a positively worded questionnaire. *Assessment & Evaluation in Higher Education*, *35*, 113-130.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph Supplement*, *17*(4, Pt. 2).

Shamay-Tsoory, S. G., Aharon-Peretz, J., & Perry, D. (2009). Two systems for empathy: A double dissociation between emotional and cognitive empathy in inferior frontal gyrus versus ventromedial prefrontal lesions. *Brain*, *132*, 617-627.

Siu, A. M., & Shek, D. T. (2005). Validation of the Interpersonal Reactivity Index in a Chinese context. *Research on Social Work Practice*, *15*, 118-126.

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*, 62-71.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245-251.

Strathearn, L., Fonagy, P., Amico, J., & Montague, P. R. (2009). Adult attachment predicts maternal brain and oxytocin response to infant cues. *Neuropsychopharmacology*, *34*, 2655-2666.

Tellegen, A., & Atkinson, G. (1974). Openness to absorbing and self-altering experiences ("absorption"), a trait related to hypnotic susceptibility. *Journal of Abnormal Psychology*, *83*, 268-277.

Thimm, J. C., Jordan, S., & Bach, B. (2016). The Personality Inventory for DSM-5 Short Form (PID-5-SF): Psychometric properties and association with big five traits and pathological beliefs in a Norwegian population. *BMC Psychology*, *4*, 61. doi:10.1186/s40359-016-0169-5

Todd, A. R., Forstmann, M., Burgmer, P., Brooks, A. W., & Galinsky, A. D. (2015). Anxious and egocentric: How specific emotions influence perspective taking. *Journal of Experimental Psychology: General*, *144*, 374-391.

Vachon, D. D., & Lynam, D. R. (2016). Fixing the problem with empathy: Development and validation of the affective and cognitive measure of empathy. *Assessment*, *23*, 135-149.

Vachon, D. D., Lynam, D. R., & Johnson, J. A. (2014). The (non) relation between empathy and aggression: Surprising results from a meta-analysis. *Psychological Bulletin*, *140*, 751-773.

Watson, D., & Clark, L. A. (1984). Negative affectivity: The disposition to experience aversive emotional states. *Psychological Bulletin*, *96*, 465-490.

Wickramasekera, I. E., & Szlyk, J. P. (2003). Could empathy be a predictor of hypnotic ability? *International Journal of Clinical and Experimental Hypnosis*, *51*, 390-399.

Wood, J. L., James, M., & Ciardha, C. Ó. (2014). "I know how they must feel": Empathy and judging defendants. *European Journal of Psychology Applied to Legal Context*, *6*, 37-43.

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, *28*, 189-194.

Young, S., Sedgwick, O., Perkins, D., Lister, H., Southgate, K., Das, M., . . . Gudjonsson, G. H. (2015). Measuring victim empathy among mentally disordered offenders: Validating VERA-2. *Journal of Psychiatric Research*, *60*, 156-162.