Empathy Measurement in Autistic and Non-autistic Adults:

A COSMIN Systematic Literature Review

Jessica L. Harrison, Charlotte L. Brownlow, Michael J. Ireland, and Adina M. Piovesana

School of Psychology and Counselling, University of Southern Queensland

[Accepted in Assessment, 2020]

Author Note

Jessica L. Harrison https://orcid.org/0000-0001-7888-7906

Charlotte L. Brownlow https://orcid.org/0000-0001-8497-5021

Michael J. Ireland https://orcid.org/0000-0001-6064-3575

Adina M. Piovesana https://orcid.org/0000-0001-8728-711X

We have no conflict of interest to disclose.

Correspondence concerning this article should be addressed to Jessica L. Harrison, School of Psychology and Counselling, University of Southern Queensland, 11 Salisbury Road, Ipswich, Queensland, 4305, Australia. Email: jessica.harrison@usq.edu.au

Abstract

Empathy is essential for social functioning and is relevant to a host of clinical conditions. This COSMIN review evaluated the empirical support for empathy self-report measures used with autistic and non-autistic adults. Given autism is characterized by social differences, it is the subject of a substantial proportion of empathy research. Therefore, this review uses autism as a lens through which to scrutinize the psychometric quality of empathy measures. Of the 19 measures identified, five demonstrated ‘High Quality’ evidence for ‘Insufficient’ properties and cannot be recommended. The remaining 14 had noteworthy gaps in evidence and require further evaluation before use with either group. Without tests of measurement invariance or differential item functioning, the extent to which observed group differences represent actual trait differences remains unknown. Using autism as a test case highlights an alarming tendency for empathy measures to be used to characterize, and potentially malign vulnerable populations before sufficient validation.

Keywords: autism, empathy, empathy quotient, empathy measurement, self-report, COSMIN, validity

Empathy Measurement in Autistic and Non-autistic Adults:

A COSMIN Systematic Literature Review

Empathy is a fundamental ability that is necessary for interpersonal communication (Bora et al., 2007; Chapman, 2016) and healthy relationships (Grühn et al., 2008). Historically, the correct referents for the term *empathy* have been the subject of disagreement among researchers (Kosonogov, 2014), with little consensus as to whether it is a cognitive or affective construct (Baron-Cohen & Wheelwright, 2004; Davis, 1980). More recently, however, empathy has been defined as a construct comprised of both components, cognitive and affective (Baron-Cohen & Wheelwright, 2004; Davis, 1980; Dziobek et al., 2008). The affective component has been defined as the ability to detect or identify another's emotional state and to vicariously experience that emotional state with them (Erol et al., 2017). The cognitive component has been defined as the ability to understand another's thoughts, feelings, and subjective experiences (Erol et al., 2017). The current review follows this approach to empathy by defining it as a two-component construct comprised of both cognitive and affective features.

Given there are minor inconsistencies in how each component is defined across studies, it is important to provide the concrete explicit definitions that form the basis of the current study. While previous researchers suggest the identification of another's emotions is part of the affective component, we expect it would also rely on cognitive and perceptual processes. Therefore, in the interest of a stronger delineation between the components, we conceptualize the affective features as involving the ability to experience a sense of the emotional states of others, while we conceptualize the cognitive feature as involving the ability to identify others' feelings and to understand their associated thoughts, feelings, and behaviors. Therefore, for the current review, the term *empathy* is reserved for referents involving both the cognitive and affective components. For this reason, the concept of 'theory of mind’, which constitutes only the cognitive feature is not considered to represent empathy in its entirety. There has been some tendency to conflate theory of mind with empathy given similarities in conceptualisation with what we believe are the cognitive features of empathy. Space does not permit us to detail these issues here except to stipulate our position that empathy is a broader construct involving both cognitive and affective features.

Importance of Empathy

Empathy is essential to communication (Bora et al., 2007), maintaining positive relationships (Grühn et al., 2008), and broader social functioning (Bailey et al., 2008; Blanke et al., 2016). It is of little surprise, therefore, that empathy has been linked to a range of positive outcomes including higher subjective well-being (Blanke et al., 2016), life satisfaction, and positive affect (Grühn et al., 2008). Empathy has also been linked to altruistic motivations to help others (Stocks et al., 2009, such as a willingness to offer money and time to those in need (Pavey et al., 2012). Therefore, it is essential for everyday social outcomes and interpersonal functioning as well as having professional consequences, especially for those in the caring professions, such as doctors and psychologists (Burks & Kobus, 2012; Elliott et al., 2018). Empathic individuals may be perceived as more caring and understanding (Grühn et al., 2008), which could further facilitate positive interactions across a variety of contexts. Given the wide-ranging effects of empathy, valid empathy evaluation is applicable to a range of contexts including treatment planning for those with interpersonal difficulties, aptitude testing for the caring professions, and evaluating empathy interventions designed to promote prosocial behavior. Empathy differences and deficits are also implicated in a host of clinical conditions with over 30 references to empathy in the DSM-IV (American Psychiatric Association [APA], 1994). Unfortunately, space does not permit us to discuss empathy regarding all of these diagnostic populations so the current analysis pays specific attention to autistic populations. We have opted to use empathy in autism as a focusing device for this review because autistic individuals are more likely to be exposed to empathy assessment and have decisions about themselves made on the basis of this assessment, and secondarily, it is more likely that the necessary volume of evidence will be available to support the planned analyses. Given the role of empathy in social functioning, it is not surprising that autism, a condition characterized by differences in social functioning, is subject to a substantial proportion of empathy research and are a focal population used in the development and validation of empathy instruments.

Autism

Autism, or Autism Spectrum Disorder, is a neurodevelopmental condition characterized by social difficulties and difficulties with repetitive behaviors (APA, 2013). The social difficulties include "deficits" in social-emotional reciprocity, nonverbal communication (e.g., gestures), and "deficits in developing, maintaining and understanding relationships" (APA, 2013, p. 50). The interaction of these characteristics with a social environment designed for the predominant neurotype (PNT; i.e. non-autistic[1] people; see Beardon, 2008) leads to an increased risk of social isolation, bullying, and abuse (Jawaid et al., 2012). It is, therefore, important to examine the individual and environmental factors that underpin these social difficulties to ensure they are properly understood and mitigated. This includes an appropriate examination of empathy in autistic[1] adults.

Empathy in Autism

Autistic people have been uniformly characterized as lacking in empathy (Bird & Viding, 2014; Cascia & Barr, 2017; Klapwijk et al., 2016), with some researchers reporting this as a central, defining feature of the condition (Fletcher-Watson & Bird, 2020; Kajganich, 2013; Rogers et al., 2007). The veracity of these conclusions, however, is entirely dependent

[1] Identity first language is chosen for this paper to reflect the understanding of autism as an integral part of an individual's identity and allow an individual to choose their identity and reclaim the label as a reflection of cultural pride (see American Psychological Association, 2020; Beardon, 2008).

on their being derived from psychometrically sound measurement. Knowledge derived from unsound measures is not only inconclusive but potentially harmful as it may reinforce damaging stereotypes and fuel stigma (see Fletcher-Watson & Bird, 2020).

Given empathy deficits are putatively a central feature of autism, empathy evaluation often forms a part of the diagnostic process. Indeed, cases have been identified where autistic adults have reported being denied an autism diagnosis solely based on them demonstrating empathy (Harrison et al., 2019). Autistic adults have also reported experiencing stigma, discrimination, and restricted career opportunities due to the assumption that they have an empathy deficit (Harrison et al., 2019). Given the impact of assumed empathy deficits, it is important to ensure it is well-founded with strong empirical support. Such empirical support must be derived from evidence gathered using psychometrically sound empathy measures.

Empathy deficits in autism appear to be empirically supported, with research regularly reporting autistic samples to score significantly lower than PNTs on measures of overall empathic abilities. For example, in a pilot study of the Empathy Quotient (EQ; Baron-Cohen & Wheelwright, 2004), autistic adults scored significantly lower than PNTs, with a small effect size. These findings have been replicated with modified forms of the EQ with children (Auyeung et al., 2009) and adolescents (Johnson et al, 2009), again with small effect sizes. Researchers who have evaluated cognitive and affective empathy separately have typically found autistic adults to be similar to PNTs on affective empathy though lower on cognitive empathy (Klapwijk et al., 2016; Montgomery et al., 2016; Rogers et al., 2007; Rueda et al., 2015; Trimmer et al., 2017). Studies by Rueda et al. (2015) and Rogers et al. (2007) present illustrative examples of this distinction. Rueda et al. (2015) used two subscales of the Interpersonal Reactivity Index (IRI) and compared empathy between diagnosed autistic adults and PNT adults. Autistic adults scored significantly lower than PNTs on a cognitive subscale called Perspective Taking (moderate effect), but similar to PNTs on an affective subscale called Empathic Concern (non-significant). Similarly, in Rogers et al.’s (2007) study, autistic adults scored significantly lower than healthy PNTs on both cognitive subscales of the IRI: Perspective Taking (small effect) and Fantasy (small effect). The results on affective empathy were mixed. While no differences emerged on Empathic Concern, autistic adults scored significantly higher than PNTs on Personal Distress (small effect), indicating greater empathy. These results suggest a more complex state of empathy differences between autistic and PNT samples, and one that is not reducible to the crude notion of empathy deficits among autistic samples.

The overall pattern of research findings to date shows a general trend for lower empathy scores in autistic individuals, primarily in cognitive empathy with small effect sizes. This reduced empathy has been associated with difficulties with social reciprocity, communication, and developing and maintaining relationships (Cascia & Barr, 2017). Despite research suggesting these differences are not reducible to mere empathy deficits, researchers and practitioners have developed several empathy interventions to improve social outcomes for autistic children (Argott et al., 2017), adolescents (Goldingay et al., 2013), and adults (Koch et al., 2015; Koehne et al., 2016). Clinical interventions such as these are costly and require significant time investment on the part of the client and practitioner (see Koch et al., 2015; Koehne et al., 2016). Given the costs, it is important to ensure that safe and empirically validated empathy interventions are provided when needed; that is, only when the individual has low empathy. It is unethical to impose ill-informed interventions upon vulnerable populations, especially as it means diverting resources from other means of support for them. To prevent this, practitioners should administer a psychometrically-sound measure to evaluate individual empathy before intervening. However, we cannot currently be certain whether, and to what extent, there is an empathy deficit in autism, given even a cursory review of the literature reveals a possible cause of spurious empathy deficits, that is, empathy measurement. Indeed, a recent editorial implicated flawed empathy measurement and theory as contributors to the "mischaracterisation of autistic people as lacking empathy" (Fletcher-Watson & Bird, 2020, p. 3).

Empathy Measurement

Prevailing understandings of empathy are derived primarily through the administration of empathy self-report measures, however, some studies have employed additional forms of measurement such as fMRI (Hoffmann et al., 2016; Klapwijk et al., 2016), behavioral observation (Melchers et al., 2015; Robinson & Robert, 2016), physiological response (Dethier et al., 2015; Holzhauer et al., 2017), and observer/parent-report (Cascia & Barr, 2017). While these other forms of measurement may appear to be more objective because they are less prone to a social desirability bias, they come with other limitations. For example, behavioral measures have demonstrated poor convergent validity (Melchers et al., 2015), observer/parent reports are prone to observer bias and the misinterpretation of the qualitatively different behavior present in autism (APA, 2013), and physiological measures have potential confounds caused by sensory issues and the autistic individual's sensory response to the measuring apparatus. It is for these reasons that the current review focused on empathy self-report measures, which are also the most commonly used approach to measuring empathy.

Empathy Self-report Measures

Research on empathy has been dominated by two self-report measures: the EQ (Baron-Cohen & Wheelwright, 2004) and the IRI (Davis, 1980, 1983). The EQ was piloted with autistic and PNT samples (Baron-Cohen & Wheelwright, 2004) and considers empathy to be comprised of interrelated cognitive and affective components that are so closely linked they cannot be separated (Baron-Cohen & Wheelwright, 2004). The EQ is, therefore, a unidimensional measure that is reported to capture both components of empathy together. It has been adapted into several additional forms to improve its reliability and validity (e.g. Lawrence et al., 2004) and to test various models of empathy (e.g. Muncer & Ling, 2006). An example item from the original EQ is “I can pick up quickly if someone says one thing but means another”. In contrast to the EQ, the IRI was developed only with PNTs (Davis, 1983) and is comprised of interrelated cognitive and affective components that are evaluated separately. The IRI contains items such as “I sometimes try to understand my friends better by imagining how things look from their perspective” (Davis, 1980, 1983). The IRI has also been adapted into a Brief IRI to address issues with the factor structure and reliability of the original measure (Ingoglia et al., 2016).

While empathy research is dominated by the EQ and IRI, numerous other self-report measures have been used. These include the Basic Empathy Scale (Jolliffe & Farrington, 2006), Hogan Empathy Scale (Hogan, 1969), Toronto Empathy Questionnaire (Spreng et al., 2009), Empathy Components Questionnaire (ECQ; Batchelder et al., 2017), and the Questionnaire for Cognitive and Affective Empathy (QCAE; Reniers et al., 2011). The Basic Empathy Scale, ECQ, and QCAE utilize a similar structure to the IRI with interrelated cognitive and affective components, which are evaluated separately (Batchelder et al., 2017; Davis, 1980; Jolliffe & Farrington, 2006; Reniers et al., 2011). In contrast, the authors of the Toronto Empathy Questionnaire and Hogan Empathy Scale acknowledge both cognitive and affective components of empathy, however, the focus of these measures is more on the affective, and cognitive components, respectively (Hogan, 1969; Spreng et al., 2009).

Self-Report Measurement Issues. While most empathy self-report measures have a low required reading level (e.g. see EQ items; Baron-Cohen & Wheelwright, 2004), they contain many culturally specific non-literal phrases. For example, the popular EQ and IRI have numerous items with non-literal phrases such as “pick up”, “how things look”, and “touched by things”. The research team estimates the proportion of non-literal items to be

40% and 43% for the EQ and IRI, respectively. Such phrases require the individual to not only be able to read and understand the words but to also be familiar with the non-literal phrases being used. This, along with typically "vague and imprecise" language (Fletcher-Watson & Bird, 2020, p. 4) may reduce item comprehensibility across cultures and some clinical groups. In particular, due to their difficulties interpreting and responding to non-literal language (APA, 2013; Gold et al., 2010; Martin & McDonald, 2004; Olofson et al., 2014), autistic people may have greater difficulty interpreting such items and be less likely to endorse them. This difficulty could generate a bias in the direction of poorer performance among autistic samples. Indeed, linguistic item bias, especially the use of idiomatic language, has been reported as the most common form of item bias (Rust & Golombok, 2009). Also, the difficulty experienced in trying to understand these items could produce a priming effect, whereby an autistic individual's frustration at trying to understand confusing items may make them less likely to endorse items of sensitivity and care that are indicative of empathy.

Interpretation issues, test bias, and priming may all contribute to the impaired performances reported in autistic samples. To ensure empathy self-report measures containing non-literal language are suitable for evaluating empathy in autistic individuals, comprehensibility must be established directly with autistic individuals. This can be established through evaluating content validity, and then confirmed through analysis of measurement invariance and differential item functioning. In addition to comprehensibility concerns, other aspects of content validity must be considered with autistic individuals. Given autistic people are a neurologically distinct population (Brownlow, 2010; Owren & Stenhammer, 2013), the relevance and comprehensiveness of empathy items must be established with autistic people through item development and a content validity evaluation.

Ethical Imperatives in Autism Research

Researchers must adhere to a series of ethical principles during their conduct, including the principle of validity (National Health and Medical Research Council [NHMRC], 2018; Yan & Munir, 2004). This principle requires researchers to follow empirically-sound, valid methodologies to protect participants from unnecessary risks (Yan & Munir, 2004). Of relevance here, is the risk of misleading results that could be relied upon in future research and diagnostic decisions (Yan & Munir, 2004). Such could be the case with the reported empathy deficit in autism. As discussed above, this apparent deficit could be the result of measurement and validity issues, which means the stigma, discrimination, and refusal of diagnosis reported in Harrison et al. (2019) could be unnecessary harms to autistic people. Research with highly stigmatized populations such as autistic adults needs to be conducted in an ethical and accessible manner and reported in a way that does not contribute to stigma and discrimination. While the focus of the current review is squarely on empathy measurement more generally, these concerns further spurred the research team's motivation to place autism at the center of this analysis. Potentially invalid (at best untested) measures should not be used in service of stigmatizing at-risk populations and reinforcing negative stereotypes that may or may not have any basis in reality. No doubt there are important differences in empathy between autistic people and PNTs, but these measures do not settle for unveiling differences, rather they scale and rank empathy abilities with an unambiguous interpretation of how low scores reflect empathic deficiencies. Therefore, the current review pays special attention to the complex interaction of familiar psychometric issues with existing vulnerabilities that exist for specific groups of individuals.

The Current Review

The overarching aim of this review was to evaluate the quality of empathy self-report measures to gain a clearer picture of their empirical bases. While the focus of the review is more broadly on empathy self-report measurement, paying particular attention to measurement in autism as a test case allows us to evaluate and comment on the common use of empathy instruments to characterize vulnerable populations and to determine whether sufficient evidence exists to supports this use. Therefore, evidence regarding these measures derived from both PNT and autistic adults will be scrutinized to produce a set of measurement recommendations for each population. Specifically, this review aimed to:

- evaluate evidence for the relevance and comprehensiveness of empathy self-report measures for autistic and PNT adults (content validity);
- evaluate evidence for the comprehensibility of empathy items for autistic adults (content validity);
- evaluate the evidence for measurement invariance, structural validity, internal consistency, reliability, measurement error, criterion validity, construct validity, and responsiveness of empathy self-report measures used with autistic and PNT adults.

Method

The methodology for this review was planned according to the PRISMA guidelines (Moher et al., 2009). The review protocol was preregistered with PROSPERO (registration # CRD42018089314, available at http://www.crd.york.ac.uk/PROSPERO/display_record.php?ID=CRD42018089314) and was conducted according to the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN; Mokkink et al., 2018; Prinsen et al., 2018; Terwee et al., 2018). The COSMIN methodology establishes guidelines for evaluating individual measurement studies, synthesizing results across numerous studies, and formulating measurement recommendations from the pooled results. For this review, *empathy* refers to the multidimensional construct defined earlier, and *non-clinical adults* refers to general community samples who do not have autism, Asperger's Syndrome, or other psychological conditions. This was determined by the reports of the primary researchers of each study.

Eligibility Criteria

For inclusion in this review, studies had to be available as full-text original articles published in English that did one of the following: (a) outlined the development of an empathy self-report measure for autistic or non-clinical adults; (b) outlined the development of a self-report measure for autistic or non-clinical adults which includes an empathy subscale; (c) validated or investigated the measurement properties of an empathy self-report measure for autistic or non-clinical adults; (d) validated or investigated the measurement properties of an empathy subscale for autistic or non-clinical adults; or (e) used an empathy self-report measure with autistic or non-clinical adults and provided evidence for the reliability, validity, or utility of the measure. Given the broad ways in which empathy is defined, there was no requirement for measures to subscribe to specific definitions for the cognitive and affective components of empathy, only that they purported to measure both components. These inclusion criteria privileged sensitivity over specificity in being deliberately broad to ensure the review gathered all relevant information to critically evaluate the self-report measures.

To focus the review on measures for English-speaking populations, studies were excluded if they were not published in English or used a non-English self-report measure. Further studies were excluded if they: (a) used the measure with a clinical group (other than autistic); (b) measured only one component of empathy (e.g., cognitive empathy, which could be confused with theory of mind); (c) used an empathy measure that was not self-report (e.g., physiological); or (d) did not report on some information pertaining to the reliability, validity, or utility of the empathy measure. There were no restrictions based on publication year, country, or publication status.

Information Sources

Original searches were conducted in May 2018 twice by two independent reviewers. A list of the information sources used, and their corresponding dates of coverage is provided in Table S1 (see supplementary data). Gray literature was sourced through searches using Google Scholar and ProQuest Dissertations and Theses using key terms, followed by ancestral searching of relevant article reference lists. Finally, after screening data and identifying relevant measures, the measure names were entered as search terms in new searches of the databases, Google Scholar, and test publisher websites. The search strategy for all publication databases is presented in Table S2.

Study Selection

Study selection was conducted over three stages, again, with each step conducted separately by two independent reviewers. First, two independent reviewers checked and removed duplicate articles using Endnote's duplicate search function followed by manual searches. Second, titles and abstracts were screened for relevance to the research questions. Finally, full-text articles were screened against the eligibility criteria. Articles that did not meet the eligibility criteria were removed. Discrepancies between the two reviewers were resolved at each stage through discussion, and where necessary, by an independent third reviewer.

Data Extraction

Data were extracted into Excel forms adapted from those provided by the COSMIN developers (Terwee & Prinsen, 2018), with additional forms developed to meet the specific needs of this review. The data items extracted, therefore, were largely guided by the COSMIN resources.

Risk of Bias in Individual Studies

Risk of bias for individual studies was evaluated using the COSMIN Risk of Bias Checklist (Mokkink et al., 2018), supported by the COSMIN guideline for systematic reviews of Patient-Reported Outcome Measures (Prinsen et al., 2018) and the COSMIN Methodology for Evaluating the Content Validity of Patient-Reported Outcome Measures: A Delphi Study (Terwee et al., 2018). This method allows reviewers to evaluate the risk of bias per measurement property, per measure, and per sample. Each study is rated as either ‘Very Good’, ‘Adequate’, ‘Doubtful’, or ‘Inadequate’ based on COSMIN’s criteria for each measurement property study. For example, for an internal consistency study to be rated as Very Good, the measurement study must report a Cronbach’s alpha or Omega statistic for each unidimensional (sub)scale separately and have no important flaws in the study design (Mokkink et al., 2018). Also, for a study of structural validity to be rated as Very Good, it must report a confirmatory factor analysis with a $N\ \geq 7\ (\text{number of items})\ \text{and}\ \geq 100,$ and there must be no important methodological flaws in the design. The criteria used to rate the risk of bias for each measurement property study are available in Mokkink et al (2018).

Data Synthesis

Measurement studies were separated into two groups: those with autistic samples and those with PNT samples. Data were then synthesized by pooling the results per measurement property, per measure. The Updated Criteria for Good Measurement properties (Prinsen et al., 2018) was applied to the pooled results to give each measurement property per measure an overall rating of ‘Sufficient’, ‘Indeterminate’, ‘Insufficient’, or ‘Inconsistent’. These criteria set out minimum standards for each measurement property to be rated as Sufficient. For example, to meet the criteria for Sufficient criterion validity, a measure must correlate with a gold standard at $r \geq .70$ (Prinsen et al., 2018).

Content Validity

Content validity was evaluated according to the COSMIN methodology outlined in Terwee et al. (2018). Specifically, measurement development and content validity studies were evaluated for risk of bias, then the results of these studies were pooled and rated against the 10 Criteria for Good Content Validity. These criteria allow reviewers to rate the relevance, comprehensiveness, and comprehensibility of each measure as either Sufficient, Insufficient, or Indeterminate.

Quality of Evidence

The quality of the pooled evidence was graded using COSMIN's modified version of the GRADE Approach (Prinsen et al., 2018; Schünemann et al., 2013). Evidence was rated on a 4-point scale from 'High' ("We are very confident that the true measurement property lies close to that of the estimate") to 'Very Low' ("We have very little confidence in the measurement property evidence: the true measurement property is likely to be substantially different to the estimate"; Prinsen et al., 2018, Table 1). When applying this approach, reviewers begin with the assumption that the pooled evidence is of High Quality. The rating is then downgraded with the emergence of concerns about risk of bias, as determined by the COSMIN Risk of Bias Checklist (Mokkink et al., 2018). Evidence ratings are downgraded further with the emergence of concerns about inconsistency, imprecision (i.e., small sample size), and indirectness (e.g., different sample; Prinsen et al., 2018; Schünemann et al., 2013).

Formulating Measurement Recommendations

When formulating recommendations, COSMIN recommends measures to be categorized as follows:

> (A) PROMs [patient-reported outcome measures] with evidence for 'Sufficient' content validity (any level) and at least 'Low Quality' evidence for 'Sufficient' internal consistency;
>
> (B) PROMs categorized not in A or C.

(C) PROMs with ‘High Quality’ evidence for an ‘Insufficient’ measurement property (Prinsen et al., 2018, p. 9)

Prinsen et al. (2018) outline category A measures as trustworthy and suitable for recommendation. Category B measures may be recommended but require further research, and category C measures should not be recommended. The expression “…High Quality evidence for an ‘Insufficient’ measurement property” is potentially jarring and awkward enough to create some pause for readers. The two qualifications contained in the statement refer to (1) a characterisation of the level of evidence (High Quality) and (2) the status of the measurement property itself (being Insufficient). This conclusion suggests that an instrument has an Insufficient measurement property and that the level of evidence to support this conclusion is of High Quality. The developers of this methodology encourage reviewers to also consider a measure’s feasibility (utility) when formulating recommendations.

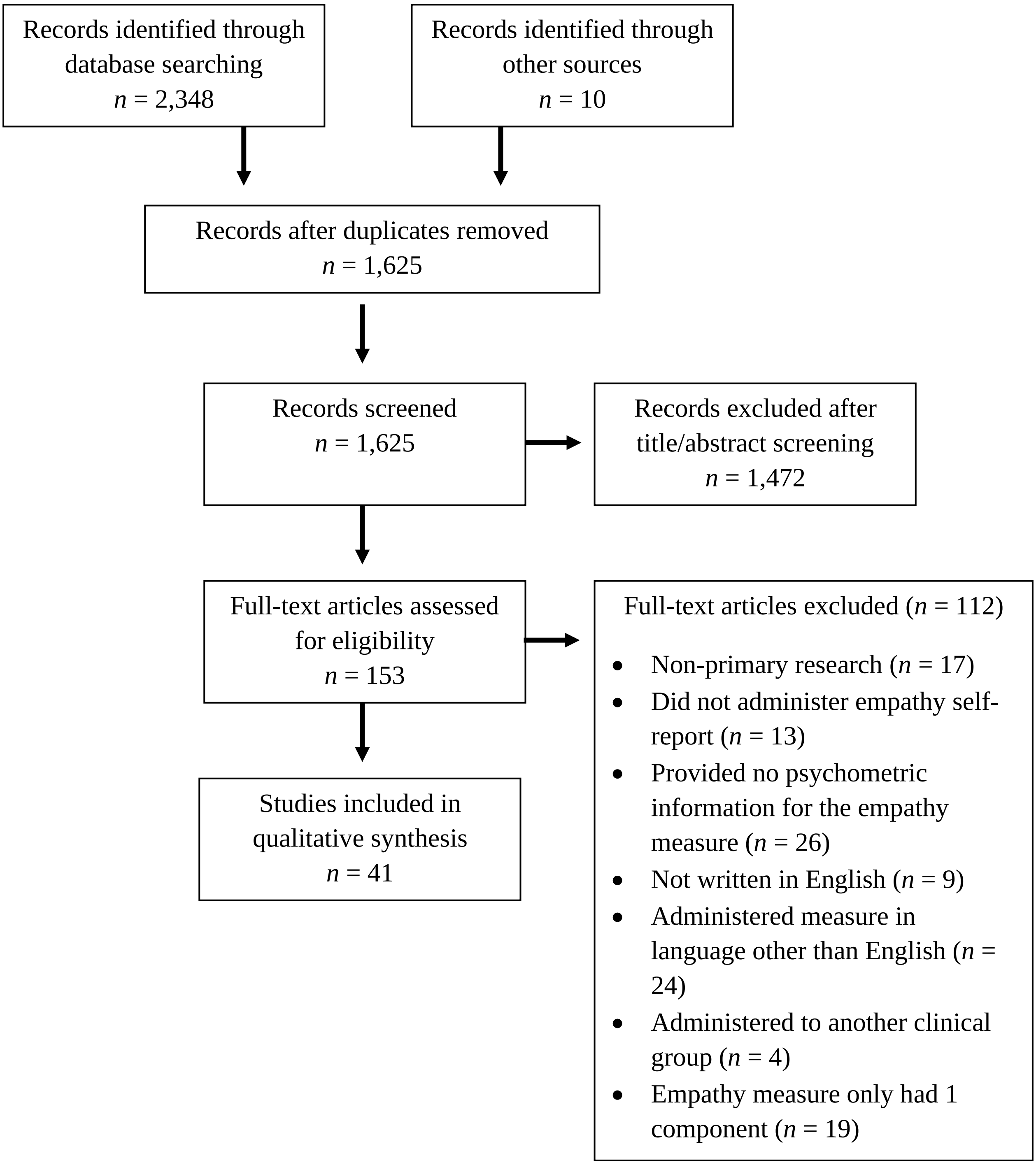
Results

Study Selection

Figure 1 displays the flow diagram of the search results through screening to the final sample of included studies. As shown, 26 articles were excluded at the full-text screening phase due to a lack of psychometric information. These articles provided no information on the empathy measure’s reliability, validity, or utility. Forty-one articles met the inclusion criteria and were included for review.

Figure 1

PRISMA flow diagram of included studies (Moher et al., 2009)


Records identified through database searching
n = 2,348
Records identified through other sources
n = 10
Records after duplicates removed
n = 1,625
Records screened
n = 1,625
Records excluded after title/abstract screening
n = 1,472
Full-text articles assessed for eligibility
n = 153
Full-text articles excluded (n = 112)
Non-primary research (n = 17)
Did not administer empathy self-report (n = 13)
Provided no psychometric information for the empathy measure (n = 26)
Not written in English (n = 9)
Administered measure in language other than English (n = 24)
Administered to another clinical group (n = 4)
Empathy measure only had 1 component (n = 19)
Studies included in qualitative synthesis
n = 41

Measurement Characteristics

A total of 19 empathy self-report measures were identified, including 10 different forms of the EQ, and two measures the research team had not yet encountered (the Perceived

Empathy Self-Efficacy Scale combined with the Perceived Social Self-Efficacy Scale; PESE/PSSE and the Just Leader). Only five of the identified measures were used with autistic samples; the IRI and four forms of the EQ. Most measures contained at least one cognitive and affective scale, with some including additional scales such as Social Skills, Fantasy, and Drive. Table 1 summarizes the measurement characteristics of each empathy self-report measure.

Table 1

Characteristics of the Included Measures

Measure	Developer/s	Target population	Scales (sub-scales)	# items	Response options	Range of scores
Empathy Quotient	Baron-Cohen & Wheelwright (2004)	"for use with adults of normal intelligence" (p. 163)[a]	Unidimensional	60	4-point	0-80
60-item Empathy Quotient – 2 subscales	Mathersul et al. (2013)	-	Cognitive, Affective	60	4-point	0-80
40-item Empathy Quotient	Allison et al. (2011)	-	Unidimensional	40	4-point	0-80
28-item Empathy Quotient – 1 factor	Muncer and Ling (2006)	-	Unidimensional	28	4-point	0-56
28-item Empathy Quotient – 3 factors	Lawrence et al. (2004)	-	Cognitive, Emotional Reactivity, Social Skills	28	4-point	0-56
26-item Empathy Quotient	Allison et al. (2011)	-	Unidimensional	26	4-point	0-52
23-item Empathy Quotient	Muncer and Ling (2006)	-	Cognitive, Emotional Reactivity, Social Skills	23	4-point	0-46
22-item Empathy Quotient – Post hoc, 3 factors	Muncer and Ling (2006)	-	Cognitive, Emotional Reactivity, Social Skills	22	4-point	0-44
22-item Empathy Quotient	Wakabayashi et al. (2006)	-	Unidimensional	22	4-point	0-44

Table 1

Characteristics of the Included Measures

Measure	Developer/s	Target population	Scales (sub-scales)	# items	Response options	Range of scores
15-item Empathy Quotient	Muncer and Ling (2006)	-	Cognitive, Emotional Reactivity, Social Skills	15	4-point	0-30
Interpersonal Reactivity Index	Davis (1980)	-	Fantasy, Perspective-taking, Empathic Concern, Personal Distress	28	5-point	0-112
Brief Interpersonal Reactivity Index	Ingoglia et al. (2016)	"General adult population" and "General adolescent population"	Fantasy, Perspective-taking, Empathic Concern, Personal Distress	16	5-point	0-64
Basic Empathy Scale	Jolliffe & Farrington, 2006	Not reported, developed with adolescent*s*[b]	Cognitive, Affective	20	5-point	20-100
Hogan Empathy Scale	Hogan, 1969	Not reported, developed with adults	-	64	True/False	-
Toronto Empathy Questionnaire	Spreng et al. (2009)	Not explicit, but "could be useful in patient populations" (p.11)	Unidimensional	16	5-point	0-80

Table 1

Characteristics of the Included Measures

Measure	Developer/s	Target population	Scales (sub-scales)	# items	Response options	Range of scores
Empathy Components Questionnaire	Batchelder et al. (2017)	“healthy and clinical populations” (p. 1)	Cognitive Ability, Cognitive Drive, Affective Ability, Affective Drive, Affective Reactivity	27	4-point	0-108
PESE & PSSE combined	Di Giunta et al. (2010)	Not reported, developed with adults	Two separate unidimensional scales combined as one	11	5-point	0-55
QCAE	Reniers et al. (2011)	Not reported, developed with university students	Perspective Taking, Emotion Contagion, Online Simulation, Peripheral Responsivity, Proximal Responsivity	31	4-point	0-124
Just Leader[c]	Graham (2017)	Leaders	Empathy	35	7-point	35-245

Note. PESE = Perceived Empathic Self-Efficacy Scale. PSSE = Perceived Social Self-Efficacy Scale. QCAE = The Questionnaire of Cognitive and Affective Empathy.

[a]Target population not reported for additional EQ forms; assume “adults of normal intelligence” as consistent with original EQ.

Table 1

Characteristics of the Included Measures

Measure	Developer/s	Target population	Scales (sub-scales)	# items	Response options	Range of scores

[b] Basic Empathy Scale was developed with adolescents, however, was also used with adults in studies included in this review.

[c] Just Leader contains an empathy scale. The number of empathy items in this scale is unreported.

As shown, the modified forms of the EQ were developed by either shortening the measure or varying the factor structure of the original measure. For example, the 60-item EQ contains the same items as the original, however, it has been separated into two unidimensional scales which provides different data for structural validity.

Sample Characteristics

Of the 41 articles identified, one included only autistic samples, 11 included both autistic and PNT samples, and 29 included only PNT samples. This equated to a total of 13 independent autistic samples (pooled *n* = 1,811) and 61 PNT samples (pooled *n* = 23,666). Table 2 presents the pooled sample size and sample description for each measure. As shown, the autistic samples were diagnosed according to varied criteria including those of the DSM-IV (APA, 1994), DSM-IV-TR (APA, 2000), DSM-5 (APA, 2013), and ICD-10 (World Health Organization, 1990). Table S3 presents the sample characteristics reported for each study separately. As shown, the gender compositions ranged from 0% to 54.9% female in the autistic samples and 0% to 100% in the PNT samples, with most autistic samples being more than 50% male and most PNT samples being more than 50% female.

Table 2

Pooled Sample Characteristics per Measure per Population

Measure	# Studies	Pooled N	Age M (SD)[a]	Gender (%F)[b]	Autism Diagnoses
Empathy Quotient					
Autistic	7	742	37.8 (12.4)	44.1%	Asperger's or "mild autism" according to DSM-IV, DSM-IV-TR, DSM-5 or ICD-10
PNT[c]	14	4,061	24.9 (10.0)	54.7%	-
60-item Empathy Quotient – 2 subscales					
Autistic	1	40	37.2 (16.2)	29.0%	"High-functioning ASD" according DSM-IV-TR
PNT	1	37	41.7 (17.2)	37.5%	-
40-item Empathy Quotient[d]					
Autistic	2	987	32.1 (11.5)	86.2%	Autism according to DSM-IV or ICD-10.

Table 2

Pooled Sample Characteristics per Measure per Population

Measure	# Studies	Pooled N	Age M (SD)[a]	Gender (%F)[b]	Autism Diagnoses
PNT	4	5,811	30.9 (11.5)	60.3%	-
28-item Empathy Quotient – 1 factor[e]					
PNT	1	362	26.3 (11.3)	53.0%	-
28-item Empathy Quotient – 3 factors					
PNT	3	254	32.3 (10.4)	54.7%	-
26-item Empathy Quotient[d]					
Autistic	1	658	30.4 (11.4)	60.7%	Autism. Criteria not specified.

Table 2

Pooled Sample Characteristics per Measure per Population

Measure	# Studies	Pooled N	Age M (SD)[a]	Gender (%F)[b]	Autism Diagnoses
PNT	1	4,719	30.4 (11.4)	60.7%	-
23-item Empathy Quotient					
PNT	1	362	26.3 (11.3)	53.0%	-
22-item EQ – post hoc 3 factors					
PNT	2	2,123	22.0 (5.6)	57.9%	-
22-item Empathy Quotient					
PNT	1	347	-	78%	
15-item Empathy Quotient					
PNT	3	796	31.1 (19.0)	59.5%	-

Table 2

Pooled Sample Characteristics per Measure per Population

Measure	# Studies	Pooled N	Age M (SD)[a]	Gender (%F)[b]	Autism Diagnoses
Interpersonal Reactivity Index					
Autistic	3	82	30.8 (13.9)	17.8%	"High-functioning" autism, Asperger's Syndrome, or pervasive developmental disorder not otherwise specified according to the DSM-IV-TR or ICD-10.
PNT	18	6,824	23.8 (9.1)	52.4%	-
Brief Interpersonal Reactivity Index[f]					
PNT	1	2,589	19.8 (3.5)	58.3%	-
Basic Empathy Scale					
PNT	3	606	20.7 (3.8)	65.7%	-
Hogan Empathy Scale					

Table 2

Pooled Sample Characteristics per Measure per Population

Measure	# Studies	Pooled N	Age M (SD)[a]	Gender (%F)[b]	Autism Diagnoses
PNT	3	484	26.8 (7.2)	62.2%	-
Toronto Empathy Questionnaire					
PNT	5	867	18.8 (2.0)	64.0%	-
Empathy Components Questionnaire					
PNT	2	312	25.3 (8.1)	58.3%	-
PESE/PSSE					
PNT	1	2,014	21.5 (20.7)	53.1%	-
QCAE					

Table 2

Pooled Sample Characteristics per Measure per Population

Measure	# Studies	Pooled N	Age M (SD)[a]	Gender (%F)[b]	Autism Diagnoses
PNT	1	640	23.7 (7.8)	67.8%	-
Just Leader					
PNT	2	1,455	-	65.9%	-

Note. # Studies = the total number of studies that have used the measure with each respective population. DSM-IV = Diagnostic and Statistical Manual of Mental Disorders, fourth edition (APA, 1994). DSM-IV-TR = DSM, fourth edition, text revision (APA, 2000). DSM 5 = DSM fifth edition (APA, 2013). ICD-10 = International Classification of Diseases, 10th edition (World Health Organization, 1990).

[a]Age mean and SD were not reported in all included studies. The pooled age mean and SD were calculated from those with reported data. [b]Gender data was not reported in all studies. The percentage of females was calculated from those with reported data. [c]Sample size was not reported in all studies evaluating the Empathy Quotient with PNT samples. [d]One study using the 40-item and 26-item Empathy Quotient reported age and gender data for all samples combined (i.e. autistic and PNT samples combined. [e]The absence of a row for autistic samples indicates a measure that had not been administered to autistic samples in any of the reviewed studies. [f]Sample data reported for adult and adolescent samples combined.

Assessment of Measurement Properties

The assessment of the nine focal measurement properties is reported for the autistic and PNT samples, separately. For conciseness, and to allow for easy comparison across measures, the results from the individual studies have been combined to produce one pooled result per measurement property, per measure, per sample. The measurement property estimates for both populations are briefly summarized in Table 3, with emboldened text indicating a measurement rating supported by High Quality pooled evidence. To identify the most supported measures for each respective population, the results are discussed for each population separately.

Table 3

Measurement Property Ratings per Measure per Sample

		Measurement property							
Measure	Sample	Content validity	Structural validity	Internal consistency	Measurement invariance	Reliability	Criterion validity	Construct validity	Responsiveness
EQ	Autistic	-[a]	?[b]	?	-	?	-	Sufficient	Insufficient
	PNT	?	?	**?**	Insufficient	?	**Insufficient**	Inconsistent	-
60-item EQ – 2 subscales	Autistic	-	-	-	-	-	-	Insufficient	-
	PNT	-	-	-	-	-	-	Insufficient	-
40-item EQ	Autistic	-	**?**	**?**	-	-	-	Sufficient	-
	PNT	-	**?**	**?**	-	-	-	Sufficient	-
28-item EQ – 1 factor[c]	PNT	-	Insufficient	**?**	-	-	-	-	-
28-item EQ – 3 factors	PNT	-	Sufficient	-	-	-	-	-	-
26-item EQ	Autistic	-	Sufficient	?	-	-	-	-	-
	PNT	-	Sufficient	?	Sufficient	-	-	-	-
23-item EQ	PNT	-	Sufficient	-	-	-	-	-	-

Table 3

Measurement Property Ratings per Measure per Sample

		Measurement property							
Measure	Sample	Content validity	Structural validity	Internal consistency	Measurement invariance	Reliability	Criterion validity	Construct validity	Responsiveness
22-item EQ – Post hoc, 3 factors	PNT	-	**Sufficient**	-	-	-	-	-	-
22-item EQ	PNT	?	Sufficient	**?**	-	-	-	**Sufficient**	-
15-item EQ	PNT	-	Sufficient	**?**	-	-	-	**Inconsistent**	-
IRI	Autistic	-	-	?	-	-	-	Inconsistent	-
	PNT	?	**?**	?	**Sufficient**	?	Insufficient	Inconsistent	-
Brief IRI	PNT	?	Sufficient	Insufficient	-	-	-	Sufficient	-
Basic Empathy Scale	PNT	-	?	?	-	-	-	Inconsistent	-
Hogan Empathy Scale	PNT	?	?	**?**	-	?	-	Inconsistent	-
TEQ	PNT	?	?	**?**	?	-	**Inconsistent**	Inconsistent	Insufficient
ECQ	PNT	?	**Insufficient**	**?**	-	-	-	**Sufficient**	-

Table 3

Measurement Property Ratings per Measure per Sample

		Measurement property							
Measure	Sample	Content validity	Structural validity	Internal consistency	Measurement invariance	Reliability	Criterion validity	Construct validity	Responsiveness
PSSE/PESE	PNT	-	**Insufficient**	**?**	Sufficient	-	-	-	-
QCAE	PNT	-	**Insufficient**	**?**	-	-	**Inconsistent**	Sufficient	-
Just Leader	PNT	?	Insufficient	?	-	-	**Insufficient**	**Sufficient**	-

Note. Emboldened text indicates a measurement property estimate supported by high-quality pooled evidence for which "We are very confident that the true measurement property lies close to that of the estimate" (Prinsen et al., 2018, Table 1).

EQ = Empathy Quotient. IRI = Interpersonal Reactivity Index. TEQ = Toronto Empathy Questionnaire. ECQ = Empathy Components Questionnaire. PESE = Perceived Empathic Self-Efficacy Scale. PSSE = Perceived Social Self-Efficacy Scale. QCAE = The Questionnaire of Cognitive and Affective Empathy.

[a] A "-" is applied to measurement property estimates for which there is no data reported in the reviewed studies. [b]A "?" is applied to measurement property estimates that are indeterminate. This occurs when the measurement property has been assessed, however, the information required to determine the sufficiency of the measurement property is not reported in the included studies. [c]The absence of a row for autistic samples indicates a measure that had not been administered to autistic samples in any of the reviewed studies.

Measurement Properties for Autistic Samples

Data were available for the psychometric properties of five measures used with autistic samples: the IRI and four forms of the EQ. There were some notable gaps in evidence, however, where none of the identified measures had data available on content validity, measurement invariance, measurement error, or criterion validity with autistic samples. When evaluating nine measurement properties over the five measures used with autistic adults, this review would, ideally, produce 45 measurement property estimates. Due to gaps in the evidence, however, this review identified data for only 13 measurement property estimates or 28.9% of those required to produce fully informed measurement decisions.

The quality of evidence for the pooled results was rated as Low to Very Low for 50% of the included studies. This represents limited confidence in the pooled estimates for these measurement properties leading to the conclusion that "the true measurement property may be…" or "is likely to be substantially different from the estimate of the measurement property" (Prinsen et al., 2018, Table 3). Also, the lack of content validity data means no determination can be made as to whether the reviewed measures are comprehensible for autistic samples. Relevance and comprehensiveness can also not be determined, and the effects of non-literal language remain unknown.

The reviewed studies reported on evidence for the structural validity, internal consistency, reliability, construct validity, and responsiveness of the empathy self-report measures used. Tables S5 to S11 present the pooled result for each measurement property across studies, the measurement property rating, and an assessment of the quality of the pooled evidence. To summarize, structural validity was Sufficient for only one measure: the 26-item EQ, with High Quality evidence. Construct validity (hypotheses testing) was rated as Sufficient for the original EQ and 40-item EQ with Moderate Quality evidence. Construct validity was rated Insufficient for the 60-item EQ with Very Low Quality evidence, and Inconsistent for the IRI with Low Quality evidence. The remaining measurement properties were either not assessed in a particular measure or rated as Indeterminate.

Due to the limited data available with autistic samples, it is worth considering the results for each study separately. These data are presented in Tables S12 and S13. These measurement property ratings are presented with the risk of bias ratings for each study. As shown, the EQ demonstrated Sufficient construct validity in five separate studies, however, the risk of bias was rated Doubtful to Inadequate indicating a ‘Very Serious’ to ‘Extremely Serious’ risk of bias (according to COSMIN criteria; Prinsen et al., 2018). Notably, the EQ also had Insufficient responsiveness, which was supported by a Very Good rated study, though it did have a very small sample size that limits the confidence one can have in the conclusions.

The measurement properties of the additional EQ forms were mostly supported by studies with Doubtful to Inadequate ratings for risk of bias. Some exceptions include the Indeterminate structural validity and internal consistency ratings of the 40-item EQ. While these results were supported by Very Good measurement studies, the studies themselves did not report the information required by COSMIN to determine the sufficiency of the respective measurement properties. The Interpersonal Reactivity Index follows the same pattern, with most studies rated as Doubtful to Inadequate for risk of bias, except for the Indeterminate internal consistency rating, which was supported by a study rated as Very Good. The Indeterminate ratings and poor risk of bias ratings limit the understandings that can be derived from the studies that included autistic samples.

Measurement Recommendations for Autistic Samples

As described in the methods section, formulating measurement recommendations involves categorising the instruments as either: (a) “recommended for use”, (b) “require further research", or (c) "should not be recommended for use" (Prinsen et al., 2018, p. 9). Table 4 presents the ratings and recommendations for measures used with autistic samples. As shown, all five measures used with autistic samples – the IRI and four forms of the EQ – require further research before they can be recommended for use.

Table 4

Measurement Recommendations for Autistic Samples

Measure	Rating	Recommendation
Empathy Quotient	B	Requires further research
60-item Empathy Quotient – 2 subscales	B	Requires further research
40-item Empathy Quotient	B	Requires further research
26-item Empathy Quotient	B	Requires further research
Interpersonal Reactivity Index	B	Requires further research

Prinsen et al. (2018) encourage reviewers to also consider a measure's feasibility, or utility when formulating recommendations. Part of the clinical utility evaluation involved calculating each measure's reading level using the Flesch Reading Ease and Flesch-Kincaid Reading Level metrics in Microsoft Word. As shown in Table 5, the clinical utility of the four EQ forms and the IRI are largely similar. The EQ forms offer a lower required reading level than the IRI, and the 26-item EQ is the shortest measure, offering greater utility. However, this form of the EQ has less empirical support than the IRI, making it difficult to select one measure as superior. Based on the COSMIN criteria, none of these measures can be recommended for clinical or general research use with autistic samples until psychometric quality has been demonstrated through targeted evaluation. Researchers and clinicians should also carefully consider these gaps in psychometric evidence when interpreting evidence from studies that have relied on these measures.

Table 5

Clinical Utility of Empathy Self-Reports

	Empathy Quotient[a]	Interpersonal Reactivity Index[b]	Basic Empathy Scale	Hogan Empathy Scale	Toronto Empathy Questionnaire	Empathy Components Questionnaire	PESE/PSSE	QCAE	Just Leader
Patient's comprehensibility									
Flesch Reading Ease	74.6	64.6	76.6	-	69.8	65.7	55.1	70.9	61.5
Flesch-Kincaid Reading Level	5.7	7.7	5.2	-	6.3	7	7.8	6.7	7.8
General	Understand CS-NLPs	Understand CS-NLPs	Understand CS-NLPs	Understand CS-NLPs	Understand CS-NLPs	Understand CS-NLPs	Understand 1 CS-NLP	Understand CS-NLPs	Understand CS-NLPs
Clinician's comprehensibility	As above	As above	As above	As Above	As above	As above	As above	As above	As above
Length of the instrument	60 items	28 items	20 items	64 items	16 items	27 items	11 items	31 items	35 items
Cost of an instrument	Free	Free	Contact authors	Purchase article	Free	Free	Free	Purchase article	Free
Mode of administration	Paper, computer	Paper, computer	Paper, computer	Paper, computer	Paper, computer	Paper, computer	Paper, computer	Paper, computer	Paper, computer

Note. PESE = Perceived Empathic Self-Efficacy Scale. PSSE = Perceived Social Self-Efficacy Scale. QCAE = The Questionnaire of Cognitive and Affective Empathy. CS-NLP = Culturally specific non-literal phrase.

[a] For conciseness, the clinical utility of the nine remaining versions of the EQ is not reported.

[b] For conciseness, the clinical utility of the Brief IRI is not reported.

Measurement Properties for PNT Samples

Data were available for the measurement properties of 19 measures used with PNT samples, including 10 forms of the EQ. Measurement error data were not reported for any of the 19 measures. When evaluating the nine measurement properties of interest across the 19 measures used with PNTs, this review would, ideally, produce 152 pooled measurement property estimates. Due to gaps in the evidence, however, this review identified data for only 59 measurement property estimates or 34.50% of those required to produce fully informed measurement decisions.

The quality of evidence, like that with the autistic samples, ranged from Very Low to High. However, unlike the evidence with autistic samples, only 27.1% of the pooled results were given Low to Very Low ratings for quality of evidence. For the remaining measurement studies (72.9%), we can be moderately to very confident that the "true measurement property is likely to be…" or is "close to that of the [pooled] estimate" (Prinsen et al., 2018, Table 3). Given the pooled evidence is of Sufficient quality, it was considered appropriate to present the results as pooled estimates according to COSMIN.

Content validity data were available for seven measures; however, all were rated as Indeterminate (see Table S4). The pooled results for the remaining measurement properties are available in Tables S5 to S11. These tables present a significant amount of information so for conciseness, this summary focuses on the results supported by High Quality evidence, that is, the results for which we can have the greatest confidence. For structural validity, the 26-item EQ was rated as Sufficient while the ECQ, PESE/PSSE, and QCAE were Insufficient. Measurement invariance was Sufficient in the IRI. Criterion validity was Insufficient for the EQ and Just Leader while both the Toronto Empathy Questionnaire and QCAE were Inconsistent. Construct validity was Sufficient in the 22-item EQ, ECQ, and Just Leader, and Inconsistent in the 15-item EQ. All remaining measurement properties were either: (a) not assessed for a measure, (b) Indeterminate, or (c) supported by lower-quality evidence.

As shown in Table 3 and the supplementary tables, five measures had High Quality pooled evidence for at least one Insufficient measurement property. Structural validity was Insufficient in the ECQ, the PESE/PSSE, and the QCAE, because these measures did not meet the minimum criteria of "CFI or TLI or comparable measure >0.95 or RMSEA <0.06 OR SRMR <0.08" (Prinsen et al., 2018, Table 1). The Just Leader had Insufficient criterion validity because it did not meet the criteria of correlating with a gold standard at $r = \geq .70$ OR AUC < .70 (Prinsen et al., 2018). Across two studies, the Just Leader was correlated with the IRI subscales, resulting in correlations ranging from $r = .17 - .48$ with IRI affective subscales and $r = .59 - .87$ with the IRI cognitive subscales (Graham, 2017). The EQ also had High Quality pooled evidence to suggest Insufficient criterion validity. Researchers correlated the EQ with the Reading the Mind in the Eyes Test ($r = .10 - .29$; Calvi, 2009; Lawrence et al., 2004), the IRI total score ($r = .40 - .67$; Calvi, 2009; Lyons et al., 2017), and the IRI subscales ($r = -.16 - .63$; Calvi, 2009; Lawrence et al., 2004). None of these correlations met the criteria of $r = \geq .70$.

Measurement Recommendations for PNT Samples

Table 6 presents recommendations for the use of each measure. As shown, not one of the 19 measures can be recommended for use without further research. Fourteen measures require further research to demonstrate psychometric quality, and five measures, those with High Quality evidence for an Insufficient property, cannot be recommended for use. Of those measures recommended for further research, the IRI and Brief IRI have the greatest number of Sufficient ratings, and thus appear to be the most promising measures.

Table 6

Measurement Recommendations for PNT Samples

Measures	Rating	Recommendation
Empathy Quotient	C	Not recommended for use
22-item Empathy Quotient[a]	B	Requires further research
Interpersonal Reactivity Index	B	Requires further research
Brief Interpersonal Reactivity Index	B	Requires further research
Basic Empathy Scale	B	Requires further research
Hogan Empathy Scale	B	Requires further research
Toronto Empathy Questionnaire	B	Requires further research
Empathy Components Questionnaire	C	Not recommended for use
PESE/PSSE	C	Not recommended for use
QCAE	C	Not recommended for use
Just Leader	C	Not recommended for use

Note. PESE = Perceived Empathic Self-Efficacy Scale. PSSE = Perceived Social Self-Efficacy Scale. QCAE = The Questionnaire of Cognitive and Affective Empathy.

[a]The eight remaining EQ versions were rated B.

When considering clinical utility (see Table 5), the Brief IRI has greater utility than the IRI, however, the IRI has greater psychometric support. Researchers should use the supplementary tables (Tables S4 to S11) plus the clinical utility data (Table 5) to make fully informed measurement decisions to meet the needs of their study.

Discussion

Research into empathy has increased significantly since 1993 (observed from “empathy” search Scopus metrics, 2019). Such research has been primarily reliant on self-report instruments, and the development and validation of these instruments have been routinely applied to samples believed to have deficient empathy such as autistic adults. Therefore, while the current review was concerned with empathy measurement in a general sense, the focus on autistic populations serves as a powerful lens through which to scrutinize the psychometric quality and common-use of these instruments. Typically, lower or ‘deficient’ empathy has been observed among autistic individuals; a group characterized by social and communication ‘impairments’. Given the reported results are often consistent with theorized deficits, it is little surprise that inadequate scrutiny has been placed on issues of bias and measurement artefacts. The theoretical claim that empathy deficits exist among autistic people precedes the development of most of these instruments and has been instrumental to the validation efforts for many of them. In turn, the apparent discovery of such deficits yielded by these same instruments is taken as confirming evidence for the theoretical claim itself. This circularity is only tenable if compelling auxiliary evidence exists to corroborate the validly of both the instruments as well as the theoretical claim.

Further motivating our concern with empathy measurement in autism is the fact that empathy measurement has crucial implications for autistic individuals and the often stigmatized, autistic community (Fletcher-Watson & Bird, 2020). The conviction that autistic individuals are deficient in empathy has spurred efforts to manipulate empathy (through psychological interventions) in this population and has also been reported to affect the diagnostic process. It is important, therefore, that conclusions regarding empathy in autism are driven by evidence gathered by empirically-sound measures. Nevertheless, this review also has important implications for PNT adults and therefore, this discussion will summarize what inferences can be supported by the existing evidence for both groups.

Gaps in Empirical Support

The scope of this review was kept deliberately broad to capture all evidence relating to empathy self-report measures used with autistic and PNT adults. Despite this, there were some considerable gaps in evidence. For the PNT samples, most measurement properties for most measures had not been evaluated at all. The gaps in evidence were even more apparent for autistic samples with no data available for content validity, measurement invariance, measurement error, or criterion validity in any of the identified measures, and only Indeterminate evidence for internal consistency and test-retest reliability. Evidence for the EQ across both groups was limited possibly by the fact that so much research effort has been directed at developing alternate forms of the measure.

Further, a total of 26 potentially relevant articles had to be excluded during the full-text screening phase because they failed to report any psychometric data for the measures used. The authors of these studies did not evaluate a single measurement property to confirm the suitability of the measure for their specific samples. This is insufficient when administering a measure within the population for which it was developed and is even more problematic when administering a measure developed with (and presumably for) PNTs to autistic samples. There is substantial evidence that autistic individuals represent a neurologically distinct group (APA, 2013; Baker, 2006), yet, none of the 19 measures in this review had established measurement invariance with autistic samples. In short, all of the identified measures, for both populations, have significant gaps in psychometric evidence and require further research and evaluation. All results derived from the use of these measures should, therefore, be interpreted with caution.

Insufficient Reporting Practices

Insufficient reporting has hampered the conclusions of this review. The pooled results for most measurement properties could only be rated as Indeterminate because the required information to determine sufficiency had not been reported. This means the significant amount of research with PNTs, for example, provides very little information on the measures' psychometric qualities. Some of the Indeterminate ratings could be due to the use of novel data analytic techniques. The Updated Criteria for Good Measurement Properties (Prinsen et al., 2018) used here, sets out the statistical data that are required to determine the sufficiency of each measurement property. Where studies have used novel data analytic techniques, they may not provide the required data set out in this criterion, thus resulting in an Indeterminate rating.

Insufficient Measurement Properties

Of the 19 empathy self-report measures reviewed, five had High Quality pooled evidence to conclude that at least one measurement property was Insufficient when used with PNT samples. These measures included the EQ, ECQ, PESE/PSSE, QCAE, and the Just Leader. Structural validity was Insufficient in the ECQ, the PESE/PSSE, and the QCAE, and criterion validity was Insufficient in the Just Leader and EQ. While these results were derived with PNT samples, the insufficient criterion validity in the EQ, which is often used with autistic samples, should be considered when choosing a measure for such samples. Given there is no data available for the EQ's criterion validity with autistic samples, researchers can only refer to that obtained with PNT samples, which resulted in the conclusion that the EQ should not be recommended for use, even for evaluation purposes. Researchers should consider this result when choosing between the EQ forms and the IRI for autistic samples and may choose to use the IRI with autistic samples, except where the participant's reading ability is below the grade 7.7 reading level required by the IRI.

Notes on the Empathy Quotient

The EQ has dominated empathy research, with 3,247 citations for the original EQ article at October 2019 (Google Scholar). As the most cited and most used empathy self-report measure, current knowledge about empathy is heavily reliant on its psychometric properties. Its popularity may also lead some researchers and practitioners to assume it is psychometrically sound. This review, however, has shown otherwise, with High Quality pooled evidence for Insufficient criterion validity in PNT samples. This result has implications for the remaining nine forms of the EQ which have not undergone criterion validity evaluation. These additional forms are restructured and short forms of the original EQ, with no alterations to item content. Given the original EQ has Insufficient criterion validity, all of the remaining EQ forms must be tested for criterion validity before use.

Limitations in Research with Autistic Samples

As discussed earlier, there is scarce research with autistic samples. The pooled sample size of only 1,811 individuals provides insufficient data with which to draw conclusions about the autistic population more generally. Also, the quality of evidence with autistic adults is significantly lower than that with PNTs. Empathy research has been conducted with comparatively less rigorous methodologies with autistic samples. Such practice breaches the ethical principles of distributive justice because it disproportionately decreases the benefits of and increases the risks of research participation to one vulnerable population (NHMRC, 2018; Yan & Munir, 2004).

Content Validity

The developers of the COSMIN methodology consider content validity to be the most important measurement property (Prinsen et al., 2018). This is because it assesses the degree to which a measure’s content reflects the construct to be measured (Prinsen et al., 2018). Unfortunately, none of the reviewed measures had evidence for content validity with autistic samples, so their relevance and comprehensiveness remain unclear. Comprehensibility is of concern here. As noted earlier, empathy self-report measures include a large proportion of items with non-literal language, yet none of the included studies reported any investigation into the effect of this language use. Without evidence to support content validity, there is no way of determining whether these measures are comprehensible or appropriate for autistic samples. In fact, given previous research has reported difficulties with non-literal language in autism (Gold et al., 2010; Martin & McDonald, 2004; Olofson et al., 2014), it is possible that these measures will not be equally comprehensible for at least those autistic adults who struggle with such language. These measures, therefore, must be subjected to content validity examination before their suitability can be known.

Measurement Invariance

Of the five measures used with autistic adults, this review identified no measurement invariance studies to establish suitability for use with this unique neurological group. It cannot determine, therefore, whether the EQ forms and IRI, developed for use with PNT samples, function equivalently with autistic samples. The distinct neurological differences between autistic and PNT adults makes it plausible that measurement variance and differential item functioning would occur. Until measurement invariance is established, using these measures to demonstrate empathy deficits in autistic individuals may be as good as using a Stroop task to examine executive functioning deficits in those with color-blindness. That is, items targeting empathy may be less endorsable for autistic respondents due to limited comprehensibility; thus, confounding results and creating spurious deficits that are merely measurement artefacts.

The lack of evidence supporting measurement invariance also has important implications for interpretations of construct validity. Construct validity can involve an assessment of known-groups validity, which is invalid in cases of measurement variance. While construct validity was rated as Sufficient for the EQ (High Quality evidence) with autistic samples, the results could be from spurious known-groups validity evidence where autistic samples scored significantly lower than PNTs as predicted. This Sufficient construct validity, therefore, needs to be interpreted with caution and with an understanding that the issue of measurement invariance has not been established, and may be confounding the results. In this way, the different measurement properties can be considered inter-dependant and therefore, establishing content validity and measurement invariance for autistic samples is more urgent.

Recommendations

Recommendations for Reporting and Interpreting Findings

A major finding of this review is the trend for researchers to fail to report sufficient psychometric information on the empathy measures used. To resolve these issues, researchers could:

- conduct some psychometric evaluation of all measures used with the samples of interest;
- report all information required to determine whether a measurement property is Sufficient or Insufficient (see the Updated Criteria for Good Measurement Properties; Prinsen et al., 2018, Table 1);
- cautiously interpret studies using empathy self-report measures with PNT adults, and with an understanding that none of the reviewed measures have sufficient empirical support; and
- interpret results with autistic samples with heightened caution, and with an understanding that most measurement properties for all measures have not been evaluated with these samples.

Recommendations for Empathy Measurement

This review has identified some significant gaps in the empirical support for empathy measures used with both autistic and PNT adults. To address these issues, researchers should:

- reconsider using the EQ, ECQ, PESE/PSSE, QCAE, and Just Leader, due to Insufficient properties;
- use the remaining measures with caution and with an understanding that they require further research;
- establish the relevance and comprehensiveness of all empathy self-report measures intended to be used with each population;

- establish the comprehensibility of all empathy self-report measures intended to be used, paying attention to issues of non-literal language with autistic samples; and
- establish measurement invariance and test for differential item functioning in all empathy self-report measures intended to be used with different populations.

Limitations of this Review

The conclusions of this review are limited to data obtained for empathy self-report measures used with autistic and PNT adults. The conclusions, therefore, cannot be generalized to other forms of empathy measurement, such as observer-report and behavioral observation. While empathy self-report measures are by far, the most popular form of empathy measurement, making the conclusions of this review widely applicable, future research could use these other forms of measurement to address the gaps in the construct validity evidence for self-report measures.

This review also inherits some limitations from its adherence to the COSMIN methodology. One major limitation impacting this review is the stringent way Indeterminate ratings are applied and then excluded from further consideration. For example, internal consistency results had to be rated as Indeterminate when structural validity was not established. While this is a reasonable requirement, it meant the Cronbach’s alphas for 18 measures were excluded from measurement evaluation. This problem was apparent across all measurement properties, with most measurement property estimates being rated as Indeterminate. A further limitation to the COSMIN approach relates to the pooling of information which results in some information attrition. To address the resulting loss of data, the results with autistic samples were discussed at the individual study level. This was not done, however, with the PNT data due to concerns of interpretability and ease of measurement comparison. The Indeterminate ratings, therefore, produced a significant loss of data in PNT samples. However, raw data is available in the supplementary material to the interested reader who may want a more fine-grained view of individual measures or measurement properties.

Conclusion

This COSMIN review critically evaluated the empirical support for empathy self-report measures used with autistic and PNT adults. The consensus of evidence suggests that none of the available measures have sufficient empirical support to be used without qualification, further investigation, and refinement. Such investigation may be futile, however, should the content of these measures be invalid. Where content validity is deemed Insufficient, the measure is unsuitable to refine; thus, researchers should instead consider the comparative benefits of developing new tools with a focus on establishing content validity as part of the development process. In addition to those measures requiring further research, five measures (EQ, ECQ, PESE/PSSE, QCAE, and the Just Leader) demonstrated Insufficient measurement properties and are, therefore, not recommended for clinical or research applications.

This review paid particular attention to the comprehensibility, relevance, and comprehensiveness of empathy self-report measures used with autistic adults. No content validity data were available. Therefore, the suitability of these measures for autistic samples remains unknown. Also, measurement invariance and differential item functioning concerning autistic samples were not assessed in any of the reviewed studies. Thus, the extent to which observed group differences represent actual trait differences also remains unknown. With insufficient psychometric support for the empathy self-report measures available, the findings of empathy deficits in autistic samples remain ambiguous and uninterpretable. The reported empathy deficit, therefore, should not be relied upon in diagnostic assessment or treatment planning until such time that this deficit is found with psychometrically-sound measures. While this review has analyzed the specific case of empathy measurement in autism as a means of scrutinizing these instruments, issues such as inattention to content validity are not exclusive to autistic populations but rather are far broader and reveal issues with the way these instruments are commonly developed and employed. Our focus on the case of autism has allowed us to highlight an alarming tendency for empathy instruments to be used to characterize, and potentially malign, vulnerable populations before their necessary validation.

References

Allison, C., Baron-Cohen, S., Wheelwright, S. J., Stone, M. H., & Muncer, S. J. (2011). Psychometric analysis of the Empathy Quotient (EQ). *Personality and Individual Differences, 51*(7), 829-835. doi:10.1016/j.paid.2011.07.005

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.).

American Psychiatric Association. (2000). *Diagnostic and statistical mannual of mental disorders* (4th ed., text rev.).

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

American Psychological Association. (2020). *Disability*. https://apastyle.apa.org/style-grammar-guidelines/bias-free-language/disability

Argott, P. J., Townsend, D. B., & Poulson, C. L. (2017). Acquisition and generalization of complex empathetic responses among children with autism. *Behavior Analysis in Practice, 10*(2), 107-117. https://doi.org/10.1007/s40617-016-0171-7

Auyeung, B., Wheelwright, S., Allison, C., Atkinson, M., Samarawickrema, N., & Baron-Cohen, S. (2009). The Children's Empathy Quotient and Systemizing Quotient: Sex differences in typical development and in Autism Spectrum Conditions. *Journal of Autism and Developmental Disorders, 39*(11), 1509-1521. https://doi.org/10.1007/s10803-009-0772-x

Bailey, P. E., Henry, J. D., & Von Hippel, W. (2008). Empathy and social functioning in late adulthood. *Aging & Mental Health, 12*(4), 499-503. http://dx.doi.org/10.1080/13607860802224243

Baker, D. K. (2006). Neurodiversity, neurological disability and the public sector: Notes on the autism spectrum. *Disability & Society, 21*(1), 15-29. https://doi.org/10.1080/09687590500373734

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*(2), 163-175. https://doi.org/10.1023/B:JADD.0000022607.19833.00

Batchelder, L., Brosnan, M., & Ashwin, C. (2017). The Development and validation of the Empathy Components Questionnaire (ECQ). *PLoS ONE, 12*(1), 1-34. https://doi.org/10.1371/journal.pone.0169185

Beardon, L. (2008). *Asperger Syndrome and perceived offending conduct: A qualitative study* [Unpublished doctoral dissertation]. Sheffield Hallam University.

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. https://doi.org/10.1016/j.neubiorev.2014.09.021

Blanke, E. S., Rauers, A., & Riediger, M. (2016). Does being empathic pay off? – Associations between performance-based measures of empathy and social adjustment in younger and older women. *Emotion, 16*(5), 671-683. http://dx.doi.org/10.1037/emo0000166

Bora, E., Gökçen, S., & Veznedaroglu, B. (2007). Empathic abilities in people with schizophrenia. *Psychiatry Research, 160*(1), 23-29. https://doi.org/10.1016/j.psychres.2007.05.017

Brownlow, C. (2010). Re-presenting autism: The construction of 'NT Syndrome'. *Journal of Medical Humanities, 31*(3), 243-255. https://doi.org/10.1007/s10912-010-9114-4

Burks, D. J., & Kobus, A. M. (2012). The legacy of altruism in health care: The promotion of empathy, prosociality and humanism. *Medical Education, 46*(3), 317-325. https://doi.org/10.1111/j.1365-2923.2011.04159.x

Calvi, J. L. (2009). *The relationship between self-report and behavioral measures of empathy* [Unpublished masters thesis]. University of North Texas.

Cascia, J., & Barr, J. J. (2017). Associations among vocabulary, executive function skills and empathy in individuals with Autism Spectrum Disorder. *Journal of Applied Research in Intellectual Disabilities, 30*(4), 627-637. https://doi.org/10.1111/jar.12257

Chapman, H. A. (2016). *Using character analysis techniques to teach cognitive empathy* [Unpublished doctoral dissertation]. Walden University.

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology, 10*, 85-104. https://fetzer.org/sites/default/files/images/stories/pdf/selfmeasures/EMPATHY-InterpersonalReactivityIndex.pdf

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, *44*(1), 113-126. https://doi.org/10.1037/0022-3514.44.1.113

Dethier, V., Bruneau, N., & Philippot, P. (2015). Attentional focus during exposure in spider phobia: The role of schematic versus non-schematic imagery. *Behaviour Research and Therapy*, *65*, 86-92. https://doi.org/10.1016/j.brat.2014.12.016

Di Giunta, L., Eisenberg, N., Kupfer, A., Steca, P., Tramontano, C., & Caprara, G. V. (2010). Assessing perceived empathic and social self-efficacy across countries. *European Journal of Psychological Assessment, 26*(2), 77-86. https://doi.org/10.1027/1015-5759/a000012.

Dziobek, I., Rogers, K., Fleck, S., Bahnemann, M., Heekeren, H. R., Wolf, O. T., & Convit, A. (2008). Dissociation of cognitive and emotional empathy in adults with Asperger Syndrome using the Multifaceted Empathy Test (MET). *Journal of Autism and Developmental Disorders, 38*(3), 464-473. https://doi.org/10.1007/s10803-007-0486-x

Elliott, R., Bohart, A. C., Watson, J. C., & Murphy, D. (2018). Therapist empathy and client outcome: An updated meta-analysis. *Psychotherapy, 55*(4), 399-410. https://doi.org/10.1037/pst0000175

Erol, A., Kirdok, A. A., Zorlu, N., Polat, S., & Mete, L. (2017). Empathy, and its relationship with cognitive and emotional functions in alcohol dependency. *Nordic Journal of Psychiatry, 71*(3), 205-209. https://doi.org/10.1080/08039488.2016.1263683

Fletcher-Watson, S., & Bird, G. (2020). Autism and empathy: What are the real links? [Editorial]. *Autism, 24*(1), 3-6. https://doi.org/10.1177/1362361319883506

Gold, R., Faust, M., & Goldstein, A. (2010). Semantic integration during metaphor comprehension in Asperger Syndrome. *Brain and Language, 113*(3), 124-134. https://doi.org/10.1016/j.bandl.2010.03.002

Goldingay, S., Stagnitti, K., Sheppard, L., McGillivray, J., McLean, B., & Pepin, G. (2013). An intervention to improve social participation for adolescents with Autism Spectrum Disorder: Pilot study. *Developmental Neurorehabilitation, 18*(2), 122-130. https://doi.org/10.3109/17518423.2013.855275

Graham, H. E. (2017). *Who is the fairest of them all: The development and validation of the Just Leader Measure* [Unpublished dissertation]. University of Texas.

Grühn, D., Rebucal, K., Diehl, M., Lumley, M., & Labouvie-Vief, G. (2008). Empathy across the adult lifespan: Longitudinal and experience-sampling findings. *Emotion, 8*(6), 753-765. https://doi.org/10.1037/a0014123

Harrison, J. L., Ireland, M., Piovesana, A. M., & Brownlow, C. (2019, June 20-23). *Empathy measurement in autism* [Paper presentaton]. Asia Pacific Autism Conference, Sentosa Island, Singapore.

Hoffmann, F., Koehne, S., Steinbeis, N., Dziobek, I., & Singer, T. (2016). Preserved self-other distinction during empathy in autism is linked to network integrity of right supramarginal gyrus. *Journal of Autism and Developmental Disorders, 46*(2), 637-648. https://doi.org/10.1007/s10803-015-2609-0

Hogan, R. (1969). Development of an empathy scale. *Journal of Consulting and Clinical Psychology, 33*(3), 307-316. http://dx.doi.org.ezproxy.usq.edu.au/10.1037/h0027580

Holzhauer, C. G., Wemm, S., & Wulfert, E. (2017). Distress tolerance and physiological reactivity to stress predict women's problematic alcohol use. *Experimental and Clinical Psychopharmacology, 25*(3), 156-165. https://doi.org/10.1037/pha0000116

Ingoglia, S., Coco, A. L., & Albiero, P. (2016). Development of a Brief Form of the Interpersonal Reactivity Index (B-IRI). *Journal of Personality Assessment, 98*(5), 461-471. https://doi.org/10.1080/00223891.2016.1149858

Jawaid, A., Riby, D. M., Owens, J., White, S. W., Tarar, T., & Schulz, P. E. (2012). 'Too withdrawn' or 'too friendly': Considering social vulnerability in two neuro-developmental disorders. *Journal of Intellectual Disability Research, 56*(4), 335-350. https://doi.org/10.1111/j.1365-2788.2011.01452.x

Johnson, S. A., Filliter, J. H., & Murphy, R. R. (2009). Discrepancies between self- and parent-perceptions of autistic traits and empathy in high functioning children and adolescents on the autism spectrum. *Journal of Autism and Developmental Disorders, 39*(12), 1706-1714. https://doi.org/10.1007/s10803-009-0809-1

Jolliffe, D., & Farrington, D. P. (2006). Deveopment and validation of the Basic Empathy Scale. *Journal of Adolescence, 29*(4), 589-611. https://doi.org/10.1016/j.adolescence.2005.08.010

Kajganich, G. (2013). *Simulation to build empathy in adolescents with Autism Spectrum Disorders: A video modeling study* [Unpublished doctoral dissertation]. University of Ottawa. https://pdfs.semanticscholar.org/e74f/006eb851e9e66829cc1f48f3711d253803b9.pdf

Kapp, S. K., Gillespie-Lynch, K., Sherman, L. E., & Hutman, T. (2013). Deficit, difference, or both? Autism and neurodiversity. *Developmental Psychology, 49*(1), 59-71. https://doi.org/10.1037/a0028353

Klapwijk, E. T., Aghajani, M., Colins, O. F., Marijnissen, G. M., Popma, A., van Lang, N. D. J., van der Wee, N. J. A., & Vermeiren, R. R. J. M. (2016). Different brain responses during empathy in Autism Spectrum Disorders versus Conduct Disorder and callous-unemotional traits. *Journal of Child Psychology and Psychiatry, 57*(6), 737-747. https://doi.org/10.1111/jcpp.12498

Koch, S. C., Mehl, L., Sobanski, E., Sieber, M., & Fuchs, T. (2015). Fixing the mirrors: A feasibility study of the effects of dance movement therapy on young adults with Autism Spectrum Disorder. *Autism: The International Journal of Research & Practice, 19*(3), 338-350. https://doi.org/10.1177/1362361314522353

Koehne, S., Behrends, A., Fairhurst, M. T., & Dziobek, I. (2016). Fostering social cognition through an imitation- and synchronization-based dance/movement intervention in adults with Autism Spectrum Disorder: A controlled proof-of-concept study. *Psychotherapy and Psychosomatics, 85*(1), 27-35. https://doi.org/10.1159/000441111

Kosonogov, V. (2014). The psychometric properties of the Russian version of the Empathy Quotient, *Psychology in Russia: State of the Art*, *7*(1), 96-104. https://doi.org/10.11621/pir.2014.0110

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*(5), 911-924. https://doi.org/10.1017/S0033291703001624

Lyons, M. T., Brewer, G., & Bethell, E. J. (2017). Sex-specific effect of recalled parenting on affective and cognitive empathy in adulthood. *Current Psychology, 36*(2), 236-241. https://doi.org/10.1007/s12144-015-9405-z

Martin, I., & McDonald, S. (2004). An exploration of causes of non-literal language problems in individuals with Asperger Syndrome. *Journal of Autism and Developmental Disorders, 34*(3), 311-328. https://doi.org/10.1023/B:JADD.0000029553.52889.15

Mathersul, D., McDonald, S., & Rushby, J. A. (2013). Understanding advanced theory of mind and empathy in high-functioning adults with Autism Spectrum Disorder. *Journal of Clinical and Experimental Neuropsychology, 35*(6), 655-668. https://doi.org/10.1080/13803395.2013.809700

Melchers, M., Montag, C., Markett, S., & Reuter, M. (2015). Assessment of empathy via self-report and behavioural paragidms: Data on convergent and discriminant validity. *Cognitive Neuropsychiatry, 20*(2), 157-171. https://doi.org/10.1080/13546805.2014.991781

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement. *PLoS Med, 6*(7). https://doi.org/10.1371/journal.pmed.1000097

Mokkink, L. B., de Vet, H. C. W., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., & Terwee, C. B. (2018). COSMIN Risk of Bias Checklist for systematic reviews of patient-reported outcome measures. *Quality of Life Research, 27*(5), 1171-1179. https://doi.org/10.1007/s11136-017-1765-4

Montgomery, C., Allison, C., Lai, M.-C., Cassidy, S., Langdon, P., & Baron-Cohen, S. (2016). Do adults with high functioning autism or Asperger Syndrome differ in empathy and emotion recognition? *Journal of Autism and Developmental Disorders, 46*(6), 1931-1940. https://doi.org/10.1007/s10803-016-2698-4

Muncer, S. J., & Ling, J. (2006). Psychometric analysis of the Empathy Quotient (EQ) scale. *Personality and Individual Differences, 40*(6), 1111-1119. https://doi.org/10.1016/j.paid.2005.09.020

National Health and Medical Research Council. (2018). *Australian Code for the Responsible Conduct of Research 2018.* https://www.nhmrc.gov.au/about-us/publications/australian-code-responsible-conduct-research-2018

Olofson, E. L., Casey, D., Oluyedun, O. A., Van Herwegen, J., Becerra, A., & Rundblad, G. (2014). Youth with Autism Spectrum Disorder comprehend lexicalized and novel primary conceptual metaphors. *Journal of Autism and Developmental Disorders, 44*(10), 2568-2583. https://doi.org/10.1007/s10803-014-2129-3

Owren, T., & Stenhammer, T. (2013). Neurodiversity: Accepting autistic difference. *Learning Disability Practice, 16*(4), 32-37. https://doi.org/10.7748/ldp2013.05.16.4.32.e681

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), 681-89. https://doi.org/10.1177/0146167211435940

Prinsen, C. A. C., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., de Vet, H. C. W., & Terwee, C. B. (2018). COSMIN guideline for systematic reviews of patient-reported outcome measures. *Quality of Life Research, 27*(5), 1147. https://doi.org/10.1007/s11136-018-1798-3

Reniers, R. L. E. P., 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*(1), 84-95. https://doi.org/10.1080/00223891.2010.528484

Robinson, A., & Robert, E. (2016). Brief report: An observational measure of empathy for autism spectrum: A preliminary study of the development and reliability of the Client Emotional Processing Scale. *Journal of Autism and Developmental Disorders, 46*(6), 2240-2250. https://doi.org/10.1007/s10803-016-2727-3

Rogers, K., Dziobek, I., Hassenstab, J., Wolf, O. T., & Convit, A. (2007). Who cares? Revisiting empathy in Asperger Syndrome. *Journal of Autism and Developmental Disorders, 37*(4), 709-715. https://doi.org/10.1007/s10803-006-0197-8

Rueda, P., Fernández-Berrocal, P., & Baron-Cohen, S. (2015). Dissociation between cognitive and affective empathy in youth with Asperger Syndrome. *European Journal of Developmental Psychology, 12*(1), 85-98. https://doi.org/10.1080/17405629.2014.950221

Rust, J., & Golombok, S. (2009). *Modern psychometrics: The science of psychological assessment* (3rd ed.). Routledge.

Schünemann, H., Brożek, J., Guyatt, G., & Oxman, A. (2013). *GRADE Handbook: Introduction to GRADE Handbook.* https://gdt.gradepro.org/app/handbook/handbook.html#h.svwngs6pm0f2

Spreng, R. N., McKinnon, M. C., Mar, R. A., & Levine, B. (2009). The Toronto Empathy Questionnaire: Scale development and initial validation of a factor-analytic solution to multiple empathy measures. *Journal of Personality Assessment, 91*(1), 62-71. https://doi.org/10.1080/00223890802484381

Stocks, E. L., Lishner, D. A., & Decker, S. K. (2009). Altruism or psychological escape: Why does empathy promote prosocial behavior? *European Journal or Social Psychology, 39*(5), 649-665. https://doi.org/10.1002/ejsp.561

Terwee, C. B. & Prinsen, C. A. C. (2018). *Help organizing your COSMIN risk of bias ratings* [Excel form]. https://www.cosmin.nl/tools/guideline-conducting-systematic-review-outcome-measures/

Terwee, C. B., Prinsen, C. A. C., Chiarotto, A., Westerman, M. J., Patrick, D. L., Alonso, J., Bouter, L. M., de Vet, H. C. W., & Mokkink, L. B. (2018). COSMIN methodology for evaluating the content validity of patient-reported outcome measures: A Delphi study. *Quality of Life Research, 27*(5), 1159-1170. https://doi.org/10.1007/s11136-018-1829-0

Trimmer, E., McDonald, S., & Rushby, J. A. (2017). Not knowing what I feel: Emotional empathy in Autism Spectrum Disorders. *Autism, 21*(4), 450-457. https://doi.org/10.1177/1362361316648520

Wakabayashi, A., Baron-Cohen, S., Wheelwright, S., Goldenfeld, N., Delaney, J., Fine, D., Smith, R., & Weil, L. (2006). Development of short forms of the Empathy Quotient (EQ-Short) and the Systemizing Quotient (SQ-Short). *Personality and Individual Differences, 41*(5), 929-940. https://doi.org/10.1016/j.paid.2006.03.017

World Health Organization. (1990). *International statistical classification of diseases and related health problems* (10th ed.). https://www.who.int/classifications/icd/icdonlineversions/en/

Yan, E. G. & Munir, K. M. (2004). Regulatory and ethical principles in research involving children and individuals with developmental disabilities. *Ethics & Behavior, 14*(1), 31-49. https://doi.org/10.1207/s15327019eb1401_3