Personality and Individual Differences 51 (2011) 829–835

Contents lists available at ScienceDirect

Personality and Individual Differences

journal homepage: www.elsevier.com/locate/paid

Psychometric analysis of the Empathy Quotient (EQ)

C. Allison [a,*], S. Baron-Cohen [a], S.J. Wheelwright [a], M.H. Stone [b], S.J. Muncer [c]

[a] Autism Research Centre, Department of Psychiatry, University of Cambridge, Douglas House, 18B Trumpington Road, Cambridge CB2 8AH, UK
[b] Psychology and Mathematics Departments, Aurora University, 347 South Gladstone Avenue, Aurora, IL 60506-4892, USA
[c] Clinical Psychology, Teesside University, Middlesbrough, TS1 3BA, UK

ARTICLE INFO

Article history:
Received 4 November 2010
Received in revised form 30 June 2011
Accepted 4 July 2011
Available online 6 August 2011

Keywords:
Autism spectrum conditions
Empathy
Rasch
Confirmatory Factor Analysis

ABSTRACT

This study assessed the dimensionality of the Empathy Quotient (EQ) using two statistical approaches: Rasch and Confirmatory Factor Analysis (CFA). Participants included $N = 658$ with an autism spectrum condition diagnosis (ASC), $N = 1375$ family members of this group, and $N = 3344$ typical controls. Data were applied to the Rasch model (Rating Scale) using WINSTEPS. The Rasch model explained 83% of the variance. Reliability estimates were greater than .90. Analysis of differential item functioning (DIF) demonstrated item invariance between the sexes. Principal Components Analysis (PCA) of the residual factor showed separation into Agree and Disagree response subgroups. CFA suggested that 26-item model with response factors had the best fit statistics (RMSEA.05, CFI .93). A shorter 15-item three-factor model had an omega (ω) of .779, suggesting a hierarchical factor of empathy underlies these sub-factors. The EQ is an appropriate measure of the construct of empathy and can be measured along a single dimension.

1. Introduction

Empathy allows us to make sense of the behaviour of others, predict what they might do next, how they feel and also feel connected to that other person, and respond appropriately to them (Wheelwright & Baron-Cohen, 2011). Empathy involves an affective and a cognitive component (Baron-Cohen & Wheelwright, 2004). The former relates to an individual having an appropriate emotional response to the mental state of another. The latter largely overlaps with the concepts of 'mindreading', or 'theory of mind': the ability to attribute mental states to others; an understanding that other people have thoughts and feelings, and that these may not be the same as your own (Baron-Cohen, 1995). Baron-Cohen and Wheelwright (2004) argue that these two components of empathy co-occur and cannot be easily disentangled.

The Empathy Quotient (EQ) (Baron-Cohen & Wheelwright, 2004) was developed as a measure of empathy because of shortcomings in existing instruments like the Interpersonal Reactivity Index (IRI) (Davis, 1980), the Questionnaire Measure of Emotional Empathy (QMEE) (Mehrabian & Epstein, 1972) and the Empathy (EM) Scale (Hogan, 1969) (see Baron-Cohen & Wheelwright, 2004; Lawrence, Shaw, Baker, Baron-Cohen, & David, 2004). The EQ is sensitive to differences in empathy in clinical and general populations; individuals with an autism spectrum condition (ASC) have reduced levels of self-reported empathy (measured by the EQ), relative to typical controls (Baron-Cohen & Wheelwright, 2004; Berthoz, Wessa, Kedia, Wicker, & Grezes, 2008; Kim & Lee, 2010; Lawrence et al., 2004; Wakabayashi et al., 2007; Wheelwright et al., 2006). The EQ shows a sex difference in empathy in the general population, females on average having higher scores than males (Baron-Cohen & Wheelwright, 2004). These findings have been replicated in cross-cultural studies in Japan (Wakabayashi et al., 2007), France (Berthoz et al., 2008) and Italy (Preti et al., 2011). A study in Korea (Kim & Lee, 2010) did not find an overall sex difference in total EQ score, an anomaly that needs to be tested further. A child parent-report version of the EQ showed a similar pattern of sex differences to that observed in adults (Auyeung et al., 2009). The EQ has clinical utility and is used as part of a screening protocol along with the Autism Spectrum Quotient (AQ) (Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001) for a clinical assessment in an adult diagnostic clinic for ASC (Baron-Cohen, Wheelwright, Robinson, & Woodbury-Smith, 2005). The EQ has convergent validity; it correlates with the 'Reading the Mind in the Eyes' Test (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001) and the Toronto Alexithymia Scale (TAS) (Lombardo et al., 2009). The EQ has been found to inversely correlate with foetal testosterone (FT) levels (Chapman et al., 2006), with single nucleotide polymorphisms (SNPs) in genes related to sex steroid hormones, neural growth, and social reward (Chakrabarti et al., 2009), and with neural activity during emotion perception in fMRI (Chakrabarti, Bullmore, & Baron-Cohen, 2006).

Lawrence et al. (2004) examined the factor structure of the EQ using Principal Components Analysis (PCA) and found a three factor solution (consisting of cognitive empathy, emotional reactivity and social skills). Berthoz et al. (2008) confirmed this structure using Confirmatory Factor Analysis (CFA). Muncer and Ling

* Corresponding author. Tel.: +44 1223 746082; fax: +44 1223 746033.
E-mail address: cla29@cam.ac.uk (C. Allison).

0191-8869/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.paid.2011.07.005

(2006) tested the unidimensionality of the EQ using CFA and found this model did not adequately fit their data. They tested other structures and confirmed that a three factor solution consisting of 15 items best fit the data. Kim and Lee (2010) confirmed this structure in their Korean sample. To date, investigation of the dimensionality of the EQ has been limited to the application of factor analysis (FA). Differences in response options (agreeing or disagreeing to selected items) may give rise to finding factors that are potentially absent or theoretically meaningless even after reverse coding of items towards the appropriate direction has occurred. Lawrence et al. (2004) point out that factor analysis on ordinal data can result in spurious factors where items load according to 'difficulty' (Gorsuch, 1974). To this end, we take a different approach at examining the dimensionality of the EQ using Rasch analysis in combination with CFA.

1.1. The Rasch model

In classical test theory (CTT), ordinal responses to questionnaire items are often treated as interval. This can lead to erroneous conclusions and inferences about the scale especially when a sum score is used to define the degree to which an individual possesses a trait or characteristic (Santor & Ramsay, 1998). Rasch (1960) developed a unique approach to psychometrics which fulfils the requirements of additive measurement (Perline, Wright, & Wainer, 1979). The principle behind Rasch analysis is as follows: 'A person having a greater ability than another person should have the greater probability of solving any item of the type in question, and similarly, one item being more difficult than another means that for any person the probability of solving the second item is the greater one' (Rasch, 1960). When participants complete a psychometric scale they provide two sources of information. One informs us how people respond to the items, (used in reliability and factor analysis studies), and the other how the participants score on the scale. This latter information is not much used in CTT. Rasch's approach uses both pieces of information when scales are analysed. The probabilistic relationship is modelled between person ability and item difficulty as a latent trait. It locates person ability and item difficulty along the same continuum in logits or log odds. The Rasch model transforms data from ordinal scores into interval level measurement with the logit.

Item difficulty is calculated using the proportion of participants who get the answer 'correct'. This is transformed into the log odds probability of getting the item correct. The ability of each participant can also be calculated, by taking the percentage of items they get correct and turning this into a probability of answering an item correctly. Rasch's theory suggests that the probability of getting an individual item correct is produced by the difference between a person's ability and the item difficulty. If a person's ability is higher than an item's difficulty, then the participant is more likely to get this correct than if it is lower than the item's difficulty. Using this information the data collected can be compared with what would be expected based on calculations of item difficulty and person ability. The closer the results are to the predicted results, the better fit the data are to the Rasch model.

Rasch analysis is designed to produce unidimensional measures when the data fit the model. Therefore, the instrument measures only one ability/personality trait/attitude. It is also designed to produce measures in which the difference between participant scores is interval scaled, making it more appropriate for statistical analysis. Rasch analysis satisfies the criteria for simultaneous conjoint measurement (Karabatsos, 2001). If the measure is unidimensional then it is reasonable to sum the item scores to produce a total score that is an adequate representation of the measured dimension. The count must be of a cohesive unit otherwise the count/measure is invalid. Rasch analysis will transform the raw counts into these cohesive units while CFA analyses the qualities of the raw ordinal (rather than interval) counts. From a Rasch perspective, items are selected to cover a wide range of the dimension, while CFA includes items that maximise reliability. Further, Rasch measures are less sensitive to directional factors (Singh, 2004) than are CFA measures.

The Rasch model has been criticised recently as not being an example of conjoint measurement (Kyngdon, 2008) (although see Michell (2008) for criticisms of Kyngdon's argument). Rasch analysis emphasises producing unidimensional measures; the main purpose of the EQ is to provide a reliable and valid measure of empathy. However, CFA is regarded as one of the most important methods for examining psychometric properties.

1.2. Aims and Objectives

We will take a pragmatic approach to examining the dimensionality of the EQ. The aims are to apply the Rasch model to a large EQ dataset to create a unidimensional measure of empathy. We then examine this model and other proposed EQ models using CFA.

2. Methods

2.1. Data source

Data included in the analysis were collected at the websites of the Autism Research Centre (ARC), University of Cambridge. Individuals can register as research volunteers and complete online questionnaires and tests. The ARC website (www.autismresearchcentre.com) recruits individuals with ASC as well as parents of children with ASC. Individuals from the general population who have an interest in taking part in research can register at www.cambridgepsychology.com. Everyone is invited to complete the Empathy Quotient (EQ). Altogether 5377 individuals completed the EQ online of which 3265 were female and 2112 were male. Within this sample, 658 individuals had a diagnosis of ASC, 1375 were family members of an individual with ASC, and 3344 had no diagnosis of ASC. The mean age of the whole sample was 30.4 years (SD = 11.4, range 16.0–78.0).

2.2. The EQ

The EQ consists of 40 statements to which participants have to indicate the degree to which they agree or disagree. There are four response options: 'strongly agree', 'slightly agree', 'slightly disagree', 'strongly disagree'. 'Definitely agree' responses score two points and 'slightly agree' responses score one point on half the items, and 'definitely disagree' responses score two points and 'slightly disagree' responses score one point on the other half. The remainder of the response options score 0. See Baron-Cohen and Wheelwright (2004) for full details.

2.3. Rasch analysis

Rasch analysis was conducted using the Rating Scale (Andersen, 1977) routine in WINSTEPS (Linacre, 2006). PROX estimation was used to converge the data with the Rasch model. The WINSTEPS reliability estimate was executed to provide an estimate of cohesion of the items (in terms of person and item reliability estimates). Item and person misfit and item Infit and Outfit statistics were examined.

Point-biserial correlations between items scores and total score were examined. It is generally agreed that these coefficient values are most acceptable for item discrimination when they occur between 0.2 and 0.8, or even closer between 0.3 and 0.7. Hence, hovering around the mean of 0.5 is considered to be ideal as a discrimination index.

The WINSTEPS principal components routine examined the item and person cohesion or lack thereof to check for unidimensionality and to assess whether a single latent trait explains the majority of the variance in the data. The ratio of variance explained by the Rasch factor to that explained by the residual factors was analysed. The step structure in item responses was examined. A reduced set of EQ items was proposed and item invariance was investigated.

2.4. Confirmatory Factor Analysis (CFA)

CFA was conducted using Amos (Arbuckle & Wothke, 1999). In all cases maximum likelihood estimation was employed, excluding cases with missing values. For each model, the chi square value and degrees of freedom, the Comparative Fit Index (CFI) and the root mean square error of approximation (RMSEA) and its confidence intervals are presented. These indices were used previously in examination of the EQ (Muncer & Ling, 2006). Browne and Cudeck's (1993) criterion for good fit was used, suggesting an RMSEA under .08 represents reasonable fit, and below .05 representing very good fit (Steiger, 1989). Higher values of the CFI are considered better with values over .9 considered acceptable.

3. Results

3.1. Rasch analysis

The data converged with the Rasch model in relatively few iterations (3 Prox and 7JMLE). A balanced spread of values led to fewer iterations required for a solution to item and person values. Few alterations to item difficulty and person ability measures were required and an acceptable Rasch model with standardized residuals of 0 was reached. This suggests that the data are already showing a balanced and cohesive order.

Item reliability for the EQ was 0.99 which is exceptional and confirms a highly cohesive set of items. See Fig. 1 for the item map, showing a graphical representation of the EQ with each item represented by its number and with the respondents grouped according to their overall EQ measure. The items are closely grouped with few 'gaps' in the sequence of items from easy to hard/rare to common. Gaps might suggest that additional items could be constructed to fill these spaces. These results show that the items occur in reasonable steps relative to each other and there is no substantial gap between them. The item distribution is well-balanced around the mean, suggesting that the EQ measures a single dimension of empathy.

In the grouping of persons, their distribution shows no clear gaps except for the few persons at either extreme which are divergent from the majority. The distributions of items and persons are quite balanced and the distributions coincide with each other; items to persons are similar by their centres and by their dispersions. This correspondence suggests that the measure is well-targeted by the items and suggests that we can be confident that the person estimate measures will be accurate (Linacre, 1994). Fig. 1 is strongly indicative of a cohesive set of items responded to by a large sample in the manner intended by the test authors.

The person reliability estimate was 0.92 which is also unusually high. Person reliability is often lower than item reliability because there is usually more erratic behaviour observed among participants than is expected among the items in a well-constructed instrument. Person misfit can also occur because of differing attributes of the respondents as well as the inclusion of a more diverse pool of individuals.

Examination of the item Outfit statistics revealed that there were no substantial deviations from expectations for most items. Two items exceeded the cut-off of 2.0 for significant outliers (items 6 and 10, see Appendix A). Item Infit statistics did not identify any deviant items. Only five items were beyond the mean of 1.06 plus the standard deviation of ±0.39 (i.e. the range between 1.45 and −0.67). These are items 6, 10, 37, 24 and 25. The same items have point-biserial correlation coefficient values outside the range of 0.3–0.7, suggesting that these items are not reliable discriminators. Examination of the person misfit statistics revealed few inconsistent responses.

The principal components analysis (PCA) of the residuals (following extraction of the Rasch component) suggests that a unidimensional solution remains appropriate for these data. The Rasch component explained 88.3% of the variance. The ratio of variance explained by the Rasch component to that explained by the residual factors was 8.6:1. Unexplained residual variance was 4.7 of 40 (11%) and was considered small enough not to influence and could be disregarded. When the PCA routine continued, the second component percentage decreased to 6%, and the third component to 4%. Taking these together, a single dimension is clearly supported. Since random variance and noise must be accounted for within the leftover variance, it is unlikely that other theoretically substantive sub-factors are operating.

Fig. 2 shows a plot of the items designated by letters (see Table 1 for how alphabetic notation in Fig. 2 relates to item number with loadings and item measure score in logits). All items are within a deviancy spread of −1 to +1 (horizontally). The vertical scale gives their calibration values. There is a general cohesiveness along both axes although some items are closer to others. Sub-groupings of items with clear separation groupings were not observed, and the item spread remained within a narrow range of deviancy horizontally, and across the calibrations shown vertically. This is further evidence to indicate that additional factors are not required.

The items for residual factor one showed separation into Agree vs. Disagree subgroups. Fig. 2 demonstrates that items 34, 14, 35, 36, 38, 11, 26, 15, 29, 22, 1, 13, 28, 4, 21, 8, 9, 40, 12 are representative of the factor and are largely items keyed in the Agree direction. The Disagree items make up the remainder.

The step structure of 0,1,2 was reasonably spaced for most items, suggesting that the response design was generally followed by the respondents. When the step structure of 0,1,2,3 was examined, the respondents did not demonstrate a gradation in responding. Fig. 3 shows a closer grouping for some of the steps than for others (indicated by an arrow at the far right in the Fig. 3). Some of these have already been identified as misfitting items, and suggests why some misfit occurred. The most extreme examples are for item 25 (between 2,3), item 23 (between 2,3), item 37 (between 1,2,3), item 6 (between 0,1,2,3), item 10 (between 1,2) and item 24 (between 0,1).

A sex difference was observed, with females (M = 0.31 SD = 0.98 logits) scoring significantly higher than males (M = −0.37 SD = 0.88; $t(4836.63) = 26$, $p < 0.0005$), $d = 0.69$. Participants with ASC scored significantly lower (M = −1.31 SD = 0.75) than controls (M = 0.23 SD = 0.88; $t(927.78) = 48.51$, $p < .0005$.), $d = 1.17$.

A shorter 26 item version of the EQ was therefore formed comprising items 1, 2, 3, 4, 7, 8, 9, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 23, 24, 26, 31, 33, 35, 36, 38, 40. This maintained the same range of item calibrations to cover the same response range as the original EQ. The spread of point-biserial correlation coefficients for indexing item discrimination was also maintained. An equal number of Agree and Disagree response keyed items were selected to avoid response bias. The items were reviewed to ensure that the content was acceptable with regards to the original theory. The significant sex difference remained, and the participants with ASC scored significantly lower than the rest of the sample.

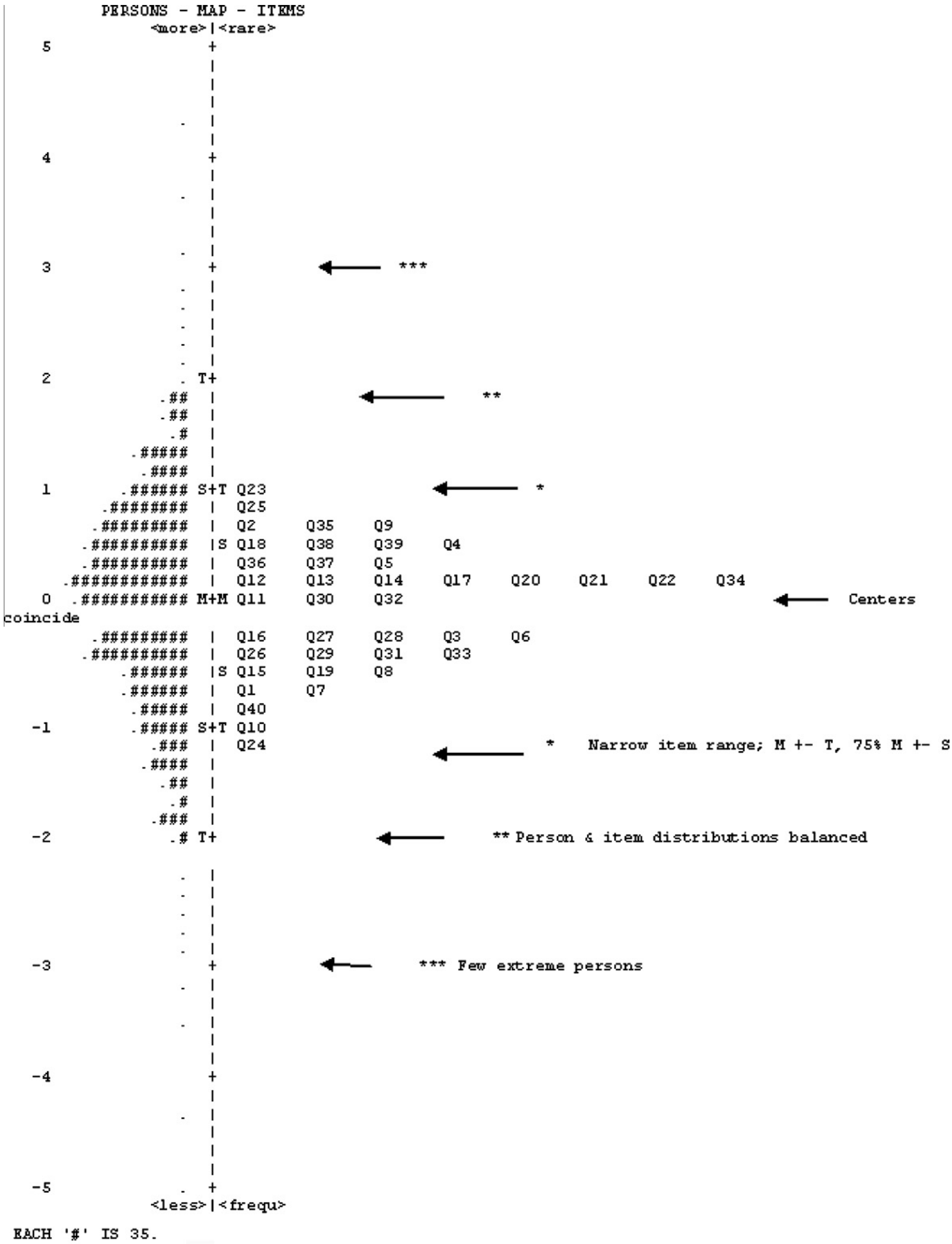

Fig. 1. Map of persons and items.

Item reliability remained very high for both samples (ASC versus not-ASC) at 0.99. Person reliability was lower but still acceptable for males ($r = 0.90$) and females ($r = 0.91$). The correlation between item calibrations for the two samples was 0.93. The average deviation of all 26 items was 0.16. The largest deviations were for item 2 (0.41) and 23 (0.34) but neither of these values was statistically significant (given a standard error of the difference of 0.452). This indicates that there were no significant differences in the item calibrations between sexes, showing that that the items were functioning similarly for both sexes.

3.2. Cfa

Results are presented in Table 2. The one factor model of the Rasch 26 item scale was not a very good fit since the CFI was low (.83). The Rasch analysis suggested that the scale may be affected by response formats which can be taken into account in CFA by either allowing for correlated error or using response formats as factors. CFA was conducted with the same 26 items modelled to include measurement factors for response format. This model had better fit statistics (CFI = 0.93; RMSEA = 0.05) than the

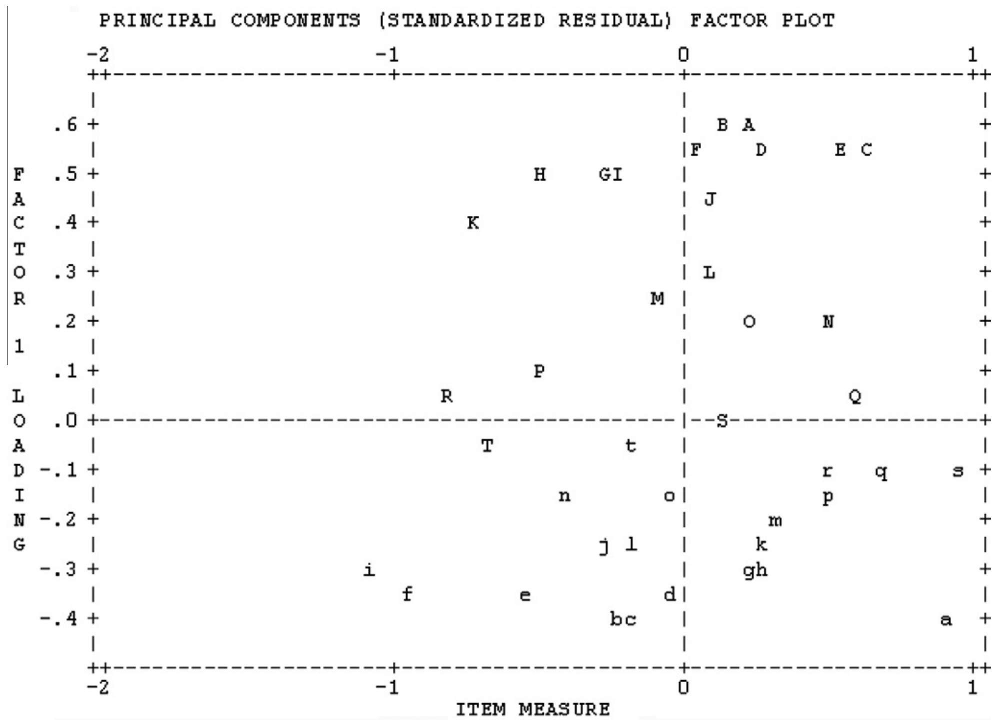

Fig. 2. Cambridge data: 5377 persons, 40 items. Principal components (standardised residual) factor plot.

Table 1
Table showing how alphabetic notation in Fig. 2 relates to item number with loadings and item measure score in logits.

Loading	Measure	Letter	Item	Loading	Measure	Letter	Item
.61	.24	A	#34	−.41	.89	a	#25
.60	.13	B	#14	−.40	−.21	b	#16
.55	.63	C	#35	−.39	−.19	c	#6
.55	.29	D	#36	−.37	−.04	d	#30
.53	.55	E	#38	−.33	−.56	e	#19
.53	.03	F	#11	−.33	−.96	f	#10
.51	−.26	G	#26	−.32	.21	g	#17
.50	−.52	H	#15	−.29	.22	h	#20
.48	−.28	I	#29	−.29	−1.09	i	#24
.45	.11	J	#22	−.27	−.28	j	#33
.41	−.72	K	#1	−.27	.26	k	#5
.32	.11	L	#13	−.23	−.18	l	#27
.24	−.11	M	#28	−.21	.30	m	#37
.21	.48	N	#4	−.17	−.40	n	#31
.21	.21	O	#21	−.17	−.03	o	#32
.09	−.49	P	#8	−.14	.48	p	#39
.06	.61	Q	#9	−.11	.67	q	#2
.03	−.83	R	#40	−.09	.50	r	#18
.02	.14	S	#12	−.08	.96	s	#23
−.07	−.67	T	#7	−.07	−.18	t	#3

one factor model, suggesting reasonable fit. Furthermore, it had a significantly better fit than the one factor model without response factors included ($\Delta\chi 2 = 5915$, df = 25 $p < .005$).

The Rasch 26-item model (empathy and response directions factors) was compared with the three factor 28-item model (Lawrence et al., 2004). This model had worse fit statistics than the Rasch model. The correlations between the three factors in the Lawrence model were high (social skills (SS)-emotional reactivity (ER) $r = 0.77$; SS-cognitive empathy (C) $r = 0.8$; ER-C $r = 0.82$) suggesting the possibility of a higher order factor of empathy.

Muncer and Ling (2006) model of the EQ as a 15-item three factor scale was tested. This model had similar fit statistics to the 26-item Rasch model with measurement factors. The 15-item three factor model was significantly improved by including response factors ($\Delta\chi 2 = 573$, df = 7, $p < .0001$). Both these scales showed a significant difference between ASC participants and controls ($t(5375) = 43.18$, $p < .0005$ for the 15-item scale, and $t(5375) = 45.2$, $p < .0005$ for the 26 item Rasch scale).

A final test of the unidimensionality of the 15-item EQ assessed whether the three factors (ER, C and SS) loaded onto a hierarchical factor. The factor loadings onto this higher factor were .73 for ER, .84 for SS and .93 for C, suggesting that a one dimensional solution is acceptable even for the 15-item version. Furthermore, omega (ω) calculated by Revelle and Zinbarg's (2009) method is 0.779, providing strong support for the view that there is a hierarchical factor of empathy underlying these subfactors.

4. Discussion

The purpose of this study was to evaluate the EQ using Rasch analysis, as a test of the potential usefulness of applying this approach to measuring the construct of empathy. Results indicated that the EQ measures a single dimension of empathy, and it is therefore acceptable to use a summed total EQ score. The data converged with the Rasch model quickly, with few iterations. This was particularly significant since this is an early indication that the EQ items are balanced and cohesive. Further, the analyses confirmed the previously reported significant difference in EQ score between the sexes and between groups (ASC versus not-ASC). The 26 item Rasch model with response factors was better than the previously suggested models that have a similar number of items.

Only five out of the 40 items were determined to be misfitting items and could be omitted from the scale, although for the sake of comparability they could also be left in as they contribute very little. Large samples can introduce more occasions for misfit than small samples when extreme deviation occurs, but this was not borne out in our large sample, indicated by only two items exceeding the significant outlier cut-off.

Item invariance was investigated for the Rasch 26-item scale. Item invariance means that the item location parameters are not

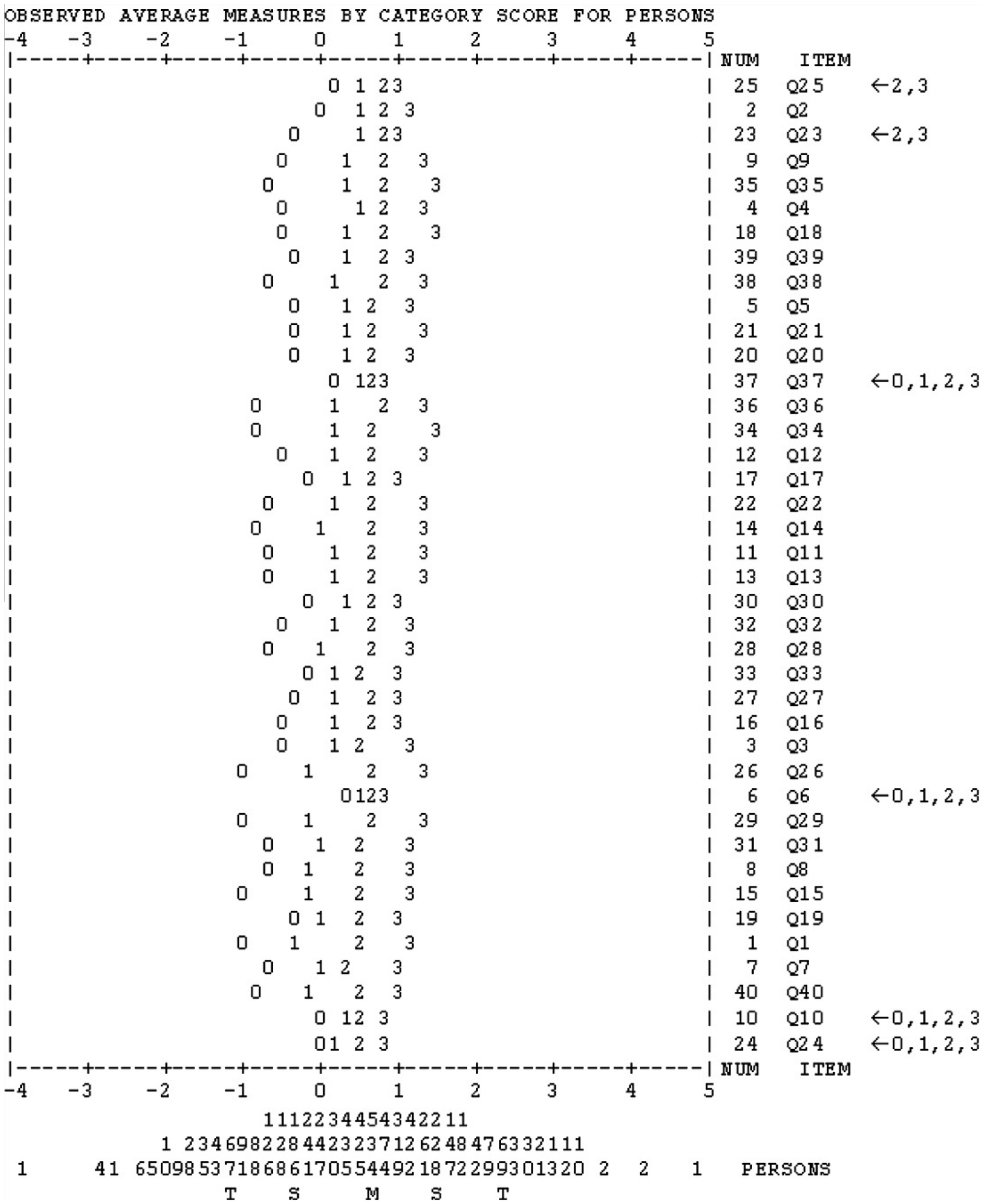

Fig. 3. 0–3 coding: 5377 persons, 40 items. Observed average measures by category score for persons.

Table 2
Table showing CFA model structures of the EQ.

Model	Items	χ^2	df	CFI	RMSEA	RMSEA 90 conf
Rasch 1 factor	26	9905	299	.83	.077	.079
Rasch 2 with measurement factors	26	3990.2	274	.93	.05	.052
Lawrence 3 factor model	28	8884	347	.88	.068	.069
Muncer and Ling 3 factor	15	1573	87	.95	.056	.059
Muncer and Ling with response factors	15	1002	80	.97	.049	.052

sample dependent. A main requirement of Rasch analysis is that items should be invariant across populations so that item parameter estimation is independent of the subgroups of individuals who complete the questionnaire. It was important to demonstrate that the EQ did have item invariance between sexes, suggesting that the items are functioning similarly for both sexes. This analysis of differential item functioning (DIF) indicates that pooling the data was justified. The 26-item EQ retains the breadth of the original 40 item EQ in terms of the level of trait captured and the ideas covered.

One way in which Rasch analysis differs from traditional approaches is that the Rasch approach analyses the residuals once the initial Rasch factor has been extracted to see what remains. In contrast, EFA produces as many factors as there are variables. Users of EFA often make much out of 'random noise' by continuing to factor in a search for 'meaning' in the data rather than have their research driven by theory. We would argue that while this approach may suggest additional factors in exploratory studies, the validity of such additional factors must rest upon further confirmation and not upon random results from exploration.

The items for residual factor one showed separation into Agree versus Disagree response subgroups. This led us to re-examine previous factor structures of the EQ scale derived through CFA. From our CFA analyses it is clear that the response format must be included as a factor as this improves all of the models.

The Rasch analysis suggests that the EQ can be considered as a unidimensional measure of empathy. CFA of this scale also supports unidimensionality as long as response direction is considered. There is some support for the view that the 15-item scale should be considered as measuring three related factors. This is evidenced by the high correlations between the factors and we therefore suggest that it is sensible to view the scale as measuring one dimension.

There are limitations to this study which must be acknowledged. First, the data were collected entirely online and therefore may contain unknown biases. Support for using the internet for data collection is provided by Gosling, Vazire, Srivastava, and John (2004), who found that data provided by internet methods are of at least as good quality as those provided by traditional methods. Further, the means and distributions of EQ data collected online and offline appear very similar (Baron-Cohen & Wheelwright, 2004; Lombardo et al., 2009). Second, the diagnoses in the group with ASC could not be verified because of the large sample size, and rely on self-report. These limitations are balanced by the strength that comes from the large sample size, which is likely to have led to robust results.

5. Conclusions

We have taken a pragmatic approach (Barrett, 2003) to measuring the construct of empathy and explored different ways of assessing this construct. This study suggests that the EQ is an appropriate measure of the construct of empathy which can be measured along a single dimension (Baron-Cohen & Wheelwright, 2004). The Rasch analysis revealed clearly that a response factor was required. The study highlights how different statistical approaches (Rasch and CFA) to measurement can be complementary, producing very similar results.

References

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42*, 69–81.

Arbuckle, J. L., & Wothke, W. (1999). *Amos users' guide, Version 4.0*. Chicago: SmallWaters Corporation.

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.

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind*. Boston: MIT Press/Bradford Books.

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.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry, 42*(2), 241–251.

Baron-Cohen, S., Wheelwright, S., Robinson, J., & Woodbury-Smith, M. (2005). The adult Asperger assessment (AAA): A diagnostic method. *Journal of Autism and Developmental Disorders, 35*(6), 807–819.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The autism-spectrum quotient (AQ): Evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders, 31*(1), 5–17.

Barrett, P. T. (2003). Beyond psychometrics: Measurement, non-quantitative structure, and applied numerics. *Journal of Managerial Psychology, 3*(18), 421–439.

Berthoz, S., Wessa, M., Kedia, G., Wicker, B., & Grezes, J. (2008). Cross-cultural validation of the empathy quotient in a French-speaking sample. *Canadian Journal of Psychiatry, 53*(7), 469–477.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In J. S. Long & K. A. Bollen (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Chakrabarti, B., Bullmore, E., & Baron-Cohen, S. (2006). Empathizing with basic emotions: Common and discrete neural substrates. *Social Neuroscience, 1*(3–4), 364–384.

Chakrabarti, B., Dudbridge, F., Kent, L., Wheelwright, S., Hill-Cawthorne, G., Allison, C., et al. (2009). Genes related to sex steroids, neural growth, and social-emotional behavior are associated with autistic traits, empathy, and Asperger syndrome. *Autism Research, 2*(3), 157–177.

Chapman, E., Baron-Cohen, S., Auyeung, B., Knickmeyer, R., Taylor, K., & Hackett, G. (2006). Fetal testosterone and empathy: Evidence from the empathy quotient (EQ) and the "Reading the Mind in the Eyes" test. *Social Neuroscience, 1*(2), 135–148.

Davis, M. (1980). A multidimensional approach to individual differences in empathy. *Catalog of Selected Documents in Psychology, 10*, 85.

Gorsuch, R. (1974). *Factor analysis*. Philadelphia: W.B. Saunders.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist, 59*(2), 93–104.

Hogan, R. (1969). Development of an empathy scale. *Journal of Consulting and Clinical Psychology, 33*, 307–316.

Karabatsos, G. (2001). The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. *Journal of Applied Measurement, 2*(4), 389–423.

Kim, J., & Lee, S. J. (2010). Reliability and validity of the korean version of the empathy quotient scale. *Psychiatry Investigation, 7*(1), 24–30.

Kyngdon, A. (2008). The Rasch model from the perspective of the representational theory of measurement. *Theory and Psychology, 18*(1), 89–109.

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), 911–919.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*(4), 328.

Linacre, J. M. (2006). *Winsteps Rasch measurement computer program*. Chicago: Winsteps.com.

Lombardo, M. V., Chakrabarti, B., Bullmore, E. T., Sadek, S. A., Pasco, G., Wheelwright, S. J., et al. (2009). Atypical neural self-representation in autism. *Brain, 133*(Pt 2), 611–624.

Mehrabian, A., & Epstein, N. (1972). A measure of emotional empathy. *Journal of Personality, 40*, 525–543.

Michell, J. (2008). Conjoint measurement and the Rasch paradox: A response to Kyngdon. *Theory and Psychology, 18*, 119–124.

Muncer, S. J., & Ling, J. (2006). Psychometric analysis of the empathy quotient (EQ) scale. *Personality and Individual Differences, 40*(6), 1111–1119.

Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement, 3*(2), 237–255.

Preti, A., Vellante, M., Baron-Cohen, S., Zucca, G., Petretto, D. R., & Masala, C. (2011). The empathy quotient: A cross-cultural comparison of the Italian version. *Cognitive Neuropsychiatry, 16*(1), 50–70.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests: Danish institute for Educational Research*. The University of Chicago Press (Expanded edition (1980)).

Revelle, W., & Zinbarg, R. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika, 74*(1), 145–154.

Santor, D. A., & Ramsay, J. O. (1998). Progress in the technology of measurement: Applications of item response models. *Psychological Assessment, 10*(4), 345–359.

Singh, J. (2004). Tackling measurement problems with item response theory: Principles, characteristics, and assessment, with an illustrative example. *Journal of Business Research, 57*(2), 184–208.

Steiger, J. H. (1989). *EzPath: Causal modelling*. Evanston, IL: Systat.

Wakabayashi, A., Baron-Cohen, S., Uchiyama, T., Yoshida, Y., Kuroda, M., & Wheelwright, S. (2007). Empathizing and systemizing in adults with and without autism spectrum conditions: Cross-cultural stability. *Journal of Autism and Developmental Disorders, 37*(10), 1823–1832.

Wheelwright, S., & Baron-Cohen, S. (2011). Systemising and Empathising. In D. A. Fein (Ed.), The neuropsychology of autism.

Wheelwright, S., Baron-Cohen, S., Goldenfeld, N., Delaney, J., Fine, D., Smith, R., et al. (2006). Predicting autism spectrum quotient (AQ) from the systemizing quotient-revised (SQ-R) and empathy quotient (EQ). *Brain Research, 1079*(1), 47–56.