Spanish Journal of Psychology (2015), **18**, e16, 1–11.

doi:10.1017/sjp.2015.19

CAMBRIDGE UNIVERSITY PRESS

Rasch Modeling and Confirmatory Factor Analysis of the Systemizing Quotient-Revised (SQ-R) Scale

Carrie Allison[1], Simon Baron-Cohen[1], Mark H Stone[2] and Steven J Muncer[3]

[1] *University of Cambridge (UK)*
[2] *Aurora University (USA)*
[3] *Teesside University (UK)*

Abstract. This study assessed the dimensionality of the Systemizing Quotient-Revised (SQ-R), a measure of how strong a person's interest is in systems, using two statistical approaches: Rasch modeling and Confirmatory Factor Analysis (CFA). Participants included $N = 675$ with an autism spectrum condition (ASC), $N = 1369$ family members of people with ASC, and $N = 2014$ typical controls. Data were applied to the Rasch model (Rating Scale) using WINSTEPS. The data fit the Rasch model quite well lending support to the idea that systemizing could be seen as unidimensional. Reliability estimates were .99 for items and .92 for persons. A CFA parceling approach confirmed that a unidimensional model fit the data. There was, however, differential functioning by sex in some of these items. An abbreviated 44-item version of the scale, consisting of items without differential item functioning by sex was developed. This shorter scale also was tested from a Rasch perspective and confirmed through CFA. All measures showed differences on total scale scores between those participants with and without ASC ($d = 0.71$, $p < .005$), and between sexes ($d = 0.53$, $p < .005$). We conclude that the SQ-R is an appropriate measure of systemizing which can be measured along a single dimension.

Received 28 November 2013; Revised 29 July 2014; Accepted 19 August 2014

Keywords: autism spectrum conditions, systemizing, rasch, confirmatory factor analysis.

Systemizing is defined as the drive to construct systems, to analyze the variables in a system, and to derive the underlying rules that govern the behavior of a system (Baron-Cohen, Richler, Bisarya, Gurunathan, & Wheelwright, 2003). This allows someone to make predictions about how a system will behave, and to control the system. The empathizing systemizing (E-S) model of sex differences suggests that males on average spontaneously systemize to a greater degree than do females (the reverse being true for empathizing) (Baron-Cohen, Wheelwright, Griffin, Lawson, & Hill, 2002). The Systemizing Quotient (SQ) (Baron-Cohen et al., 2003) was developed to measure individual differences in the drive to systemize. The aim was to tap into the following types of systems: Technical, natural, abstract, social, organisable, and motoric. The SQ has been translated into fifteen other languages, including Spanish, French, Finnish and Chinese.

The SQ was a 40 item measure with a four choice response structure ranging from *definitely agree* to *definitely disagree*. However, Wheelwright et al. (2006) argued that the items in this scale were derived "primarily from traditional male domains" (p. 48). In order to guard against circularity – that males might be more likely to score higher on the SQ because the domains are more male-typical – the final 75-item version of the SQ-R (i.e., the Revised version) includes examples of systemizing in varying degrees from everyday life, measuring social and domestic systems, as well as mechanical and abstract systems. Males on average score significantly higher on the SQ-R, relative to typical females (Baron-Cohen et al., 2003; Wheelwright et al., 2006). Individuals diagnosed with an autism spectrum condition (ASC) on average score significantly higher than people from the general population. These results have been replicated in a Japanese population (Wakabayashi et al., 2007). Child and adolescent versions of the SQ-R have been developed (Auyeung, Allison, Wheelwright, & Baron-Cohen, 2012; Auyeung et al., 2009), showing similar patterns of sex differences to those observed in adults, as well as group differences between individuals with and without ASC. SQ scores correlate with increased activation in brain regions associated with increasing and maintaining attention, and do not correlate with IQ (Billington, Baron-Cohen, & Bor, 2008; Ling, Burton, Salt, & Muncer, 2009). Nettle (2007) also found that systemizing is not explained by existing personality constructs (the Big Five factor model).

To date, few studies have examined the psychometric properties of the SQ. A Principal Components Analysis (PCA) revealed 11 factors (from 40 items) that did not correspond to factors with any obvious psychological significance (Baron-Cohen et al., 2003). Wakabayashi

Correspondence concerning this article should be addressed to Steven J Muncer. Clinical Psychology. Teesside University. Teesside (UK).
E-mail: S.Muncer@tees.ac.uk

et al. (2006) found 25 of the original 40 items loaded on to one factor and suggested that the scale measures one factor. Ling et al. (2009) conducted a confirmatory factor analysis (CFA) on 40 items, finding a one factor model was a poor fit. They also examined Wakabayashi et al.'s (2006) 25 item unidimensional model, which was also a poor fit. Further analyses led to their suggestion that the SQ is best considered as a four-factor 18-item scale, with subscales measuring topography, technicity, home improvements, and structure, with a hierarchical factor of systemizing. All analyses reported above were conducted on the original version of the SQ (40 items), which may contain items that have a male bias. These results indicate that overall the SQ's structure is not clear and that confirmatory factor analysis suggests a one dimensional structure is a poor fit to the data.

The purpose of the current study was to re-examine the dimensionality of the SQ-R (75 items) in a larger sample, using CFA, as well as a contrasting statistical approach: Rasch analysis. It was also intended to investigate the possibility of differential item functioning between male and female participants. In particular, to test for bias towards males, or indeed females, that may in appropriately influence scores on systemizing. Although Wheelwright et al. (2006) intended to develop a scale without a male item bias; so far this has not been tested.

Rasch analysis is designed to produce unidimensional measures (measuring one ability/personality trait/attitude), which is achieved when the data fit the model. A Rasch scale implies that a sum score can legitimately be used to quantify the trait. Rasch's (1960) technique creates scales that fulfill the requirements of additive measurement (Perline, Wright, & Wainer, 1979), thus treating ordinal data as interval. A person who has a greater ability than another person should have a greater probability of solving any item of that type in question. The probability of solving an easier item is greater than the probability of solving a harder item. The probabilistic relationship between person ability and item difficulty as a latent trait is modeled. Person ability (percentage of 'correct' items) and item difficulty (the proportion of participants who get the item 'correct') are located along the same continuum in logits (log odds), transforming data obtained from ordinal scores into interval level measurement. The probability of getting any item correct is produced by the difference between a person's ability and the item difficulty. If a person's ability is lower than an item's difficulty, then the participant is less likely to get this 'correct' than if it is higher than the item's difficulty. This information can be used to compare the actual data collected with expected calculations of item difficulty and person ability. The closer the actual results are to the predicted results, the better fit the data are to the Rasch model.

In the case of the items of the SQ-R we are not concerned with whether an item is correct, but on how characteristic a positive response would be for high and low 'systemetizers'. A participant with high overall scores on systemizing would be expected to definitely agree with items that are measuring this dimension. The logic and statistical procedures, however, are the same as if the SQ-R was an ability measure.

Rasch analysis will be used to try to develop a one dimensional scale to measure systemizing. It also provides a good method to examine differential item functioning by comparing item responses for male and female participants, so that we can be sure that the scale is not biased towards either sex. Lastly, confirmatory factor analysis will also be used to examine the factor structure of items that fit the Rasch model.

Method

Data source

Data were collected at the websites of the Autism Research Centre (ARC), University of Cambridge. Individuals register as research volunteers and complete online questionnaires and tests. The ARC website[1] recruits individuals with ASC as well as parents of children with ASC[2]. Everyone is invited to complete the Systemizing Quotient - Revised (SQ-R). $N = 4058$ individuals completed the SQ-R online, of which $N = 2768$ were female and $N = 1290$ were male. Within this sample, $N = 675$ individuals had a diagnosis of ASC, $N = 1369$ were family members of an individual with ASC, and $N = 2014$ had no diagnosis of ASC. The mean age of the whole sample was 38 years ($SD = 12$ years). The total sample was randomly divided into two ($N = 2029$ each), so that the analyses could be independently validated in a new sample. When the whole sample is used this is indicated in the Results.

Material

SQ-R

The SQ-R consists of 75 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

[1]www.autismresearchcentre.com

[2]Individuals from the general population can register at www.cambridgepsychology.com

https://doi.org/10.1017/sjp.2015.19 Published online by Cambridge University Press

on the other half. The remainder of the response options score 0. See Baron-Cohen et al. (2003) and Wheelwright et al. (2006) for full details.

Procedure

Rasch analysis

Initial analyses took place on Group A, who were randomly selected from the whole sample ($N = 2029$), and was then replicated in Group B ($N = 2029$). The Rating Scale (Andersen, 1977) routine in WINSTEPS (Linacre, 2006) was conducted. The WINSTEPS reliability estimate was executed to provide an estimate of cohesion of the items (person and item reliability estimates). Point-biserial correlations between items scores and total score were inspected. In Rasch analysis point-biserial correlation is used as a first check that the scoring makes sense as all items should be positively correlated (Linacre, 2006). More importantly, the fit of each of the items to the model can also be investigated. There are two fit measures, Infit and Outfit, which both reflect the ratio of the observed variance to what we would expect from the Rasch model. Infit, however, is weighted to reduce the effect of outliers or extreme responses. Item and person misfit and item Infit and Outfit statistics were examined. PROX estimation was used initially to converge the data with the Rasch model. Unconditional maximum likelihood estimation (UCON) was subsequently used.

A principal components analysis (PCA) was performed on the residuals after the Rasch measurement factor was removed. This is conducted to check for the possibility of other large components that explain the data, other than the Rasch dimension (Linacre, 1998). If there is such a component then the data do not fit a unidimensional model. The components/factors were examined to check for unidimensionality and to assess whether a single latent trait explains the majority of the variance in the data (Mavranezouli, Brazier, Young, & Barkham, 2011). The ratio of variance explained by the Rasch measurement factor to that explained by the residual factors was also analyzed (Wright & Stone, 2004).

Differential item functioning can be investigated with Rasch analysis by comparing item difficulty measures (in this case likelihood of agreement) from each sex. Items should be invariant across groups. If the 'item difficulty' estimates between the two genders shift significantly, then this suggests that the two genders are responding differently to the items and the scale overall is not measuring the same thing in both genders. In order to examine differential item functioning the Rasch analyses were subsequently carried out separately on the male sample and the female sample.

Confirmatory Factor Analysis

CFA was conducted on the full 75 item version of the SQ-R (on Group A) using Amos (Arbuckle, 2006) with maximum likelihood estimation. First a one factor solution was specified. The chi square ($\chi 2$) value and degrees of freedom, the Comparative Fit Index (CFI) and the root mean square error of approximation (RMSEA) were examined. Browne and Cudek'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.

The data were then also examined using a CFA parceling approach. A parceling approach was adopted, to reduce the number of items, as fit statistics are affected by the number of items (Kenny & McCoach, 2003), and scales with large numbers of items generally have very poor fit (Cook, Kallen, & Amtmann, 2009; Floyd & Widaman, 1995). This is because in confirmatory factor analysis the unique variances of items may be correlated because the items share a specific feature, and this becomes more likely with more items. This is particularly true also when many items are constructed from similar domains, as is the case with the SQ-R. Parcels of items are also more reliable and are also more likely to have linear relations with each other and with the proposed latent factors (Comrey, 1988). Yang, Nay, and Hoyle (2010) argue that parceling has been identified as a desirable approach when there are more than 12 items in a scale (Marsh, Hau, Balla, & Grayson, 1998) and when the items reflect a unidimensional construct (Hall, Snell, & Foust, 1999).

Results

Rasch analysis - Comparison of two sub-samples

Three UCON iterations brought the convergence to about 20% of the starting values (Group A) with four further iterations required to achieve full convergence. This small number of iterations required indicated that the data was a relatively good fit to the Rasch model, particularly given the large number of both items and participants. Item reliability was high at .99, which indicated that the item difficulties (strength of indication of systemizing) were very likely to replicate if the items were given to another sample. The person reliability was also high at .92, which indicated that the ordering of participants in terms of their systemizing was likely to replicate if they were given another parallel set of items. Both of these figures suggested that there was a viable Rasch dimension in the data.

For Group B, item reliability was 0.99 and person reliability was 0.93, following the same pattern of UCON

https://doi.org/10.1017/sjp.2015.19 Published online by Cambridge University Press

iterations, indicating the SQ-R has exceptionally high item reliability. The items coalesced around a single cohesive variable. The high reliability for persons also indicated that the sample participants responded to the SQ-R items in a cohesive manner. Item calibrations for both samples correlated at $r(75) = .99$, confirming the consistency of item calibrations across the two samples.

The PCA of the residuals following extraction of the Rasch measurement factor on Group A and Group B produced similar outcomes in each group. Items loaded onto the same components for both groups, and the loadings were of similar strength. The correlation between the two sets of item loadings from group A and group B was $r = .87$. This indicated that the two randomly derived samples were producing similar results and therefore were combined for further analysis.

Rasch analysis - Total sample statistics

Using the whole sample, item reliability was .99 and person reliability was .92, which suggested that the data were fitting the Rasch model. The point-biserial correlation were generally in the right direction with only item 3 with a negative correlation. More importantly both the Infit and Outfit statistics were indicative of a good fit of all of the items to the Rasch model. A fit value of 1 indicates perfect fit and fit values of over 1.5 are taken to indicate poor fit (Linacre, 1998). None of the items showed poor fit by that criterion, see Table 1.

The PCA of the residuals can also be used to assess the unidimensionality of the SQ-R. The eigenvalue of the residual component/factor is quite high at 5.2 and exceeds Linacre's (1998) criterion for the presence of another factor. The factor sensitivity ratio (Wright & Stone, 2004) can be used to evaluate the importance of residual factors after the Rasch measurement factor has been extracted. This is calculated by dividing the residual variance by the Rasch measurement variance. The ratio of the residual factor variance (5.2) to the Rasch measurement variance (21.6) was 0.24. This suggests that about 24% of the measure is affected by unexplained relationships between the items, which is large enough to require explanation. To put it another way about 75% of the variance in the measure can be explained by the Rasch measurement factor, but the other factor is fairly large.

The factor loadings were examined In order to try to define the other factor, besides systemizing, that might be influencing the result. An examination of the items that load positively and negatively at over .3 onto the residual first factor did not suggest that this factor had any psychological meaning, but was more likely to be method variance. All twelve of the items with strong positive loadings on the residual factor were scored in the direction of agreement, and nine of the ten items with strong negative loadings were scored in the direction of disagreement. This strongly suggests that although the SQ-R is measuring one dimension, this is influenced by the direction of scoring of the items. A very similar pattern of results was found by Allison, Baron-Cohen, Wheelwright, Stone, & Muncer (2011) in their analysis of the original Empathy Quotient scale. On that occasion they pointed out that some researchers had mistaken method variance for something more psychological.

The scores on items scored in the direction of agreement and disagreement were significantly related $r(4058) = .68$, $p < .005$, and both types of items showed large sex differences. There was, however, a significant interaction between gender and type of item $F(1, 4056) = 27.94$, $p < .005$. The difference between sexes was larger for items scored in the direction of agreement ($d = .53$), than for disagreement ($d = .47$).

Differential item functioning

Item and person reliabilities for both males and females were .99 and .92 respectively, which suggests that the scale overall works well for both sexes. One method of comparing item functioning is to conduct a t test comparison of the item measure scores from both sexes. Such a comparison has been shown to be oversensitive with reasonably large samples, so we followed Tristan's (2006) procedure for adjusting for sample size. On this basis five items showed differential item functioning. All of these had a very large difference in item 'difficulty' (0.8+ logits) between males and females. These were items 20, 32, 43, 49, and 55 (see Table 2 for the full details of SQ-R items).

For these items the gender difference was more than one would expect to be caused by differences in systemizing. Items 20 and 55 are both related to shopping and females scored more, and males less than would be expected. It is possible that these items are affected by the differences in likelihood of shopping by males and females. Item 49 was "I do not tend to remember people's birthdays (in terms of which day and month it falls)", where males scored much lower than expected. It is possible that males are responding mostly to the first part of the item, that is they do not remember birthdays, and therefore, the second part which might indicate systemizing is largely ignored. Item 32 ("I am fascinated by how machines work") and item 43 ("If there was a problem with electrical wiring in my home, I'd be able to fix it myself") have particularly low scores from females.

It is fair to say that at least four of these items are poorly constructed. Responses to the shopping and birthday items are less indicative of differences in SQ

https://doi.org/10.1017/sjp.2015.19 Published online by Cambridge University Press

Table 1. *SQ Rasch Item Statistics (75 items)*

Items	Measure	Infit MNSQ	Oufit MNSQ	CORR.
SQ1	–0.54	1.02	1.04	0.39
SQ2	–0.81	0.94	0.94	0.44
SQ3	0.19	1.43	1.62	–0.01
SQ4	–1.03	1.08	1.15	0.3
SQ5	–0.43	1.02	1.03	0.38
SQ6	–0.76	1.03	1.1	0.37
SQ7	0.79	0.89	0.81	0.49
SQ8	0.49	1.06	1.13	0.34
SQ9	0.18	0.87	0.84	0.53
SQ10	–0.5	1.06	1.11	0.36
SQ11	0.25	0.87	0.84	0.53
SQ12	0.22	0.95	0.95	0.45
SQ13	0.15	0.93	0.91	0.47
SQ14	–0.17	0.93	0.91	0.48
SQ15	–0.54	1.02	1.05	0.39
SQ16	0.46	0.8	0.74	0.59
SQ17	0.14	0.93	0.94	0.47
SQ18	0.52	0.84	0.78	0.55
SQ19	–0.1	0.89	0.86	0.52
SQ20	–0.03	1.16	1.2	0.26
SQ21	–0.26	1.01	1.01	0.41
SQ22	–0.34	1.06	1.09	0.35
SQ23	–0.42	1.14	1.2	0.25
SQ24	0.76	0.97	0.99	0.4
SQ25	0.45	1	0.99	0.4
SQ26	0	1.07	1.09	0.35
SQ27	0.19	0.79	0.75	0.61
SQ28	–0.13	1.1	1.13	0.32
SQ29	0.64	0.87	0.8	0.51
SQ30	–0.44	0.85	0.82	0.55
SQ31	–0.07	1.13	1.14	0.3
SQ32	0.2	0.8	0.76	0.59
SQ33	0.24	0.96	0.99	0.43
SQ34	0.37	1.05	1.06	0.35
SQ35	–0.69	0.95	0.96	0.45
SQ36	–0.3	1.13	1.16	0.28
SQ37	0.34	1.02	1.03	0.38
SQ38	–0.26	1.03	1.04	0.38
SQ39	0.54	1.12	1.19	0.3
SQ40	0.38	0.95	0.94	0.44
SQ41	0.17	0.84	0.82	0.55
SQ42	–0.62	1	1	0.41
SQ43	0.68	0.97	0.9	0.43
SQ44	0.12	1.13	1.18	0.29
SQ45	0.01	0.98	1.01	0.43
SQ46	0.03	0.93	0.92	0.48
SQ47	0.05	1.07	1.08	0.33
SQ48	–0.01	0.97	0.95	0.44
SQ49	0.05	1.28	1.37	0.15
SQ50	0.1	0.9	0.89	0.49
SQ51	–0.06	1.07	1.07	0.36
SQ52	–0.24	1.02	1.05	0.39
SQ53	–0.71	0.9	0.89	0.49
SQ54	–0.13	1.02	1.02	0.41
SQ55	–0.34	1.12	1.14	0.3

https://doi.org/10.1017/sjp.2015.19 Published online by Cambridge University Press

Table 1. *(Continued)*

Items	Measure	Infit MNSQ	Oufit MNSQ	CORR.
SQ56	0.06	1.02	1.03	0.39
SQ57	0.12	1.14	1.2	0.28
SQ58	–0.1	1	1	0.41
SQ59	0.21	1.11	1.13	0.3
SQ60	–0.08	0.83	0.8	0.58
SQ61	–0.22	1.08	1.11	0.33
SQ62	–0.2	1.07	1.1	0.31
SQ63	0.33	1.02	1.02	0.37
SQ64	0.25	0.93	0.92	0.47
SQ65	–0.25	1.06	1.06	0.35
SQ66	0.08	0.88	0.84	0.53
SQ67	–0.2	1.12	1.15	0.29
SQ68	0.12	1.04	1.04	0.39
SQ69	0.96	0.94	0.84	0.43
SQ70	0.42	0.89	0.87	0.51
SQ71	0.15	1.06	1.07	0.36
SQ72	–0.65	0.98	0.97	0.42
SQ73	0.06	1	1.01	0.41
SQ74	0.25	0.91	0.87	0.5
SQ75	–0.1	1.05	1.07	0.38
M	**0**	**1**	**1.01**	
SD	**0.4**	**0.11**	**0.15**	

and more related to other sex differences. The item about electricity asks not just about interest but also about a specific ability, which is likely to be more prevalent in males. Given the relatively small number of items that show severe differential item functioning, however, it is probably not necessary to omit these items from the scale. For the current data removal of these items reduced the effect size of gender on SQ total from $d = .57$ to $d = .56$. Participants with a diagnosis of Asperger's syndrome showed a similar pattern but with larger effect sizes; for the full 75 items the effect size was $d = .69$ and for the reduced version $d = .70$.

So far we have taken a relatively conservative view of differential item functioning by selecting items based on the modified t test procedure. It is possible to take a stricter definition and this is often done when it is important to rule out bias. For example, Scheuneman and Subhiyah (1998) used an item measure difference of greater than 0.5 logits to detect bias in a medical certification test given by the National Board of Medical Examiners. For the present data, we also took a stricter criterion of a logit difference of greater than 0.4, as suggesting differential item functioning between males and females. The number of items which were now omitted is 31. There are 44 items still remaining, and 23 of these are scored in the direction of disagreement. These 44 items are indicated by an underlined item number in Table 2. There is, however, still a large and significant sex difference in the expected direction $t(4056) = 16.97$, $p < .005$, with a slightly reduced effect size $d = .53$. There is also a large and significant difference for those with a diagnosis of Asperger's syndrome $t(4056) = 22.83$, $p < .005$, with a larger effect size of d = .71. It is clear that there is a substantial difference between males and females on systemizing, and also between Asperger's sufferers and non-sufferers, that cannot be ascribed to bias in the items of the Systemizing Quotient.

Confirmatory Factor Analysis

It was useful to examine the data from the two created samples separately as if modifications of the proposed structure were required then these could be validated on the second sample. Fit statistics from AMOS (Arbuckle, 2006) for Group A are presented in Table 3. The one factor solution for all 75 items was not a good fit according to the χ^2 value or the CFI, despite a reasonable RMSEA value. This is not surprising, given the problems in using CFA at item level on large tests. The items with the most problematic loadings onto a single factor were 3 (–.09) and 49 (.05). Fifty-eight of the items had loadings onto the latent factor that were greater than .3.

There are a number of different possible approaches to parceling (Landis, Beal, & Tesulk, 2000). These include parceling by content in which items are parceled according to their item content, parceling based on an

https://doi.org/10.1017/sjp.2015.19 Published online by Cambridge University Press

Table 2. *The SQ-R 75 items with indication of Agreement or Disagreement scoring procedure*

Item	Item content
1A	I find it very easy to use train timetables, even if this involves several connections.
2A	I like music or book shops because they are clearly organized.
3D	I would not enjoy organizing events e.g. fundraising evenings, fetes, conferences.
4A	When I read something, I always notice whether it is grammatically correct.
5A	I find myself categorizing people into types (in my own mind).
6D	I find it difficult to read and understand maps.
7A	When I look at a mountain, I think about how precisely it was formed.
8D	I am not interested in the details of exchange rates, interest rates, stocks and shares.
9A	If I were buying a car, I would want to obtain specific information about its engine capacity.
10D	I find it difficult to learn how to program video recorders.
11A	When I like something I like to collect a lot of different examples of that type of object, so I can see how they differ from each other.
12A	When I learn a language, I become intrigued by its grammatical rules.
13A	I like to know how committees are structured in terms of who the different committee members represent or what their functions are.
14A	If I had a collection (e.g. CDs, coins, stamps), it would be highly organized.
15D	I find it difficult to understand instruction manuals for putting appliances together.
16A	When I look at a building, I am curious about the precise way it was constructed.
17D	I am not interested in understanding how wireless communication works (e.g. mobile phones).
18A	When travelling by train, I often wonder exactly how the rail networks are coordinated.
19A	I enjoy looking through catalogues of products to see the details of each product and how it compares to others.
20A	Whenever I run out of something at home, I always add it to a shopping list.
21A	I know, with reasonable accuracy, how much money has come in and gone out of my bank account this month.
22D	When I was young I did not enjoy collecting sets of things e.g. stickers, football cards etc.
23A	I am interested in my family tree and in understanding how everyone is related to each other in the family.
24D	When I learn about historical events, I do not focus on exact dates.
25A	I find it easy to grasp exactly how odds work in betting.
26D	I do not enjoy games that involve a high degree of strategy (e.g. chess, Risk, Games Workshop).
27A	When I learn about a new category I like to go into detail to understand the small differences between different members of that category.
28D	I do not find it distressing if people who live with me upset my routines.
29A	When I look at an animal, I like to know the precise species it belongs to.
30A	I can remember large amounts of information about a topic that interests me e.g. flags of the world, airline logos.
31D	At home, I do not carefully file all important documents e.g. guarantees, insurance policies
32A	I am fascinated by how machines work.
33D	When I look at a piece of furniture, I do not notice the details of how it was constructed.
34D	I know very little about the different stages of the legislation process in my country.
35D	I do not tend to watch science documentaries on television or read articles about science and nature.
36A	If someone stops to ask me the way, I'd be able to give directions to any part of my home town.
37D	When I look at a painting, I do not usually think about the technique involved in making it.
38A	I prefer social interactions that are structured around a clear activity, e.g. a hobby.
39D	I do not always check off receipts etc. against my bank statement.
40D	I am not interested in how the government is organized into different ministries and departments.
41A	I am interested in knowing the path a river takes from its source to the sea.
42A	I have a large collection e.g. of books, CDs, videos etc.
43A	If there was a problem with the electrical wiring in my home, I'd be able to fix it myself.
44D	My clothes are not carefully organized into different types in my wardrobe.
45D	I rarely read articles or WebPages about new technology.
46A	I can easily visualize how the motorways in my region link up.
47D	When an election is being held, I am not interested in the results for each constituency.
48D	I do not particularly enjoy learning about facts and figures in history.
49D	I do not tend to remember people's birthdays (in terms of which day and month this falls).
50A	When I am walking in the country, I am curious about how the various kinds of trees differ.
51D	I find it difficult to understand information the bank sends me on different investment and saving systems.
52D	If I were buying a camera, I would not look carefully into the quality of the lens.

https://doi.org/10.1017/sjp.2015.19 Published online by Cambridge University Press

Table 2. *(Continued)*

Item	Item content
53A	If I were buying a computer, I would want to know exact details about its hard drive capacity and processor speed.
54D	I do not read legal documents very carefully.
55A	When I get to the checkout at a supermarket I pack different categories of goods into separate bags.
56D	I do not follow any particular system when I'm cleaning at home.
57D	I do not enjoy in-depth political discussions.
58D	I am not very meticulous when I carry out D.I.Y or home improvements.
59D	I would not enjoy planning a business from scratch to completion.
60A	If I were buying a stereo, I would want to know about its precise technical features.
61A	I tend to keep things that other people might throw away, in case they might be useful for something in the future.
62A	I avoid situations which I cannot control.
63D	I do not care to know the names of the plants I see.
64D	When I hear the weather forecast, I am not very interested in the meteorological patterns.
65D	It does not bother me if things in the house are not in their proper place.
66A	In maths, I am intrigued by the rules and patterns governing numbers.
67D	I find it difficult to learn my way around a new city.
68A	I could list my favorite 10 books, recalling titles and authors' names from memory.
69A	When I read the newspaper, I am drawn to tables of information, such as football league scores or stock market indices.
70D	When I'm in a plane, I do not think about the aerodynamics.
71D	I do not keep careful records of my household bills.
72A	When I have a lot of shopping to do, I like to plan which shops I am going to visit and in what order.
73D	When I cook, I do not think about exactly how different methods and ingredients contribute to the final product.
74A	When I listen to a piece of music, I always notice the way it's structured.
75A	I could generate a list of my favorite 10 songs from memory, including the title and the artist's name who performed each song.

exploratory factor analysis in which items are parceled according to factor loadings, and random parceling. As the Rasch analysis suggested that a one factor solution was possible, a random parceling approach was taken. In this approach it is assumed that each of the items is an indicator of the factor and therefore any combination of items should produce parcels which will be a good fit to a one factor solution. It was decided to create 5 parcels, which insured an equivalent number of items in each parcel, and the number of parcels would not be too large. When the 75 items were randomly parceled into five groups, the CFI suggested a good fit (see Table 3). The correlations between parcels ranged between.71 to .78. The RMSEA value of .11, however, was still higher than would be expected for a well-fitting model. The modification indices revealed that the model could be improved by allowing correlated error between parcels 2 and 4. This resulted in the CFI increasing to .99, with the RMSEA now at .05, indicating excellent model fit to the data.

Table 3. *CFA of the SQ – Group A*

Models	Items	χ^2	*df*	CFI	RMSEA
One factor	75	26322	2700	.49	.066
One factor/5 parcels	75	134	5	.98	.11

This model with correlated error between the same parcels was then tested on the data from Sample B. The pattern of results was very similar in Group B. For the 75 item SQ-R with parceling, the CFI was .99 and the RMSEA was .044, when correlated error was allowed between parcels 2 and 4. When the factor loadings, covariance of the error term between parcels 2 and 4, and the variance of the latent variable were constrained to be the same for both groups of participants, the fit of the constrained model was not significantly different to the model in which these parameters could be independently estimated ($\Delta\chi^2(6) = 8.93$, $p > .05$). This indicated that the proposed structure of the items and their loadings are relatively consistent across samples.

Aluja and Blanch (2004) have argued that a Single Factor item parceling method works well. In this procedure parcels are formed based on the factor loadings when a single factor is extracted. Parcels are formed by selecting items with high and low loadings onto the factor alternately. This parceling method was used to select five different parcels and the fit also suggested that SQ could be seen as one factor ($\chi^2(4) = 41.2$; CFI =.99; RMSEA =.048).

Parceling was also tested on the 44 items without differential item functioning between the sexes. These items

https://doi.org/10.1017/sjp.2015.19 Published online by Cambridge University Press

were randomly split into four parcels to maintain equivalent numbers between parcels, and the fit statistics were acceptable ($\chi^2(2) = 30.09$, $p < .05$; CFI =.997; RMSEA = .06). The factor loadings for all four parcels were also greater than .82. This pattern of results increases support for the proposition that the SQ is measuring a single dimension and should be treated as such.

Discussion

This study examined the psychometric properties of the SQ-R, and explored in particular whether the latent construct of systemizing is unidimensional. Results from the Rasch modeling and CFA with parcels indicated that the SQ-R measures a single dimension of systemizing, and it is therefore appropriate to use a summed SQ-R score to describe the extent to which an individual possesses a drive to systemize.

From the Rasch analysis, high item calibration reliability values were found for both male and female participants. Some items, however, showed differential item functioning by sex. When these items were omitted, however, the overall sex difference was very similar suggesting that differences cannot be ascribed to the biased items and that any differences are a consequence of sex differences in systemizing. It is possible, therefore, either to use the full 75 item version or to use the shorter 44 item version which removes items that show differential item functioning. The SQ-R in either the 75 or 44 item version, also shows differences in the expected direction between those with and without ASC.

The PCA on the residuals remaining from the Rasch analysis suggested that participants are affected by the direction of the items, as the first factor located Agree and Disagree items on opposite ends of this factor. This appeared to be the main defining element of this factor. It should be noted that such measurement factors are quite common. Allison et al. (2011) previously argued that measurement factors on the Empathy Quotient scale had been misinterpreted as being meaningful psychological factors.

This is not to say that it would be impossible to find subfactors that might be argued to exist. For example, an examination of the modification indices deriving from the CFA of the 75 item structure, suggested things could be improved by allowing correlated error between some items. For example, correlated error between item 20 *Whenever I run out of something at home I always add it to a shopping list* and item 72 *When I have a lot of shopping to do I like to plan what shops I am going to visit and in what order,* would improve the fit. However, this seems to be based merely on the subject of the item, 'shopping' and not a psychological factor of any interest. The pattern of modifications reveals a number of these item content clusters. We would argue that they have no psychological relevance and are in fact examples of bloated specifics, to which both Exploratory and Confirmatory Factor Analysis are particularly susceptible. We would argue that these are spurious results based on overlap of item content, and agree with Nunnally and Bernstein (1994, p.316) that

> "…spurious results may lead to inappropriate criticism of sound scales or, what is basically the same thing, lead an investigator to falsely believe that the scale that he or she has developed is appropriately multidimensional when in fact it is not."

A parceling approach was an efficient way of reducing the number of items on the SQ-R, of overcoming the problem of ordinal data and producing data that is closer to a normal distribution. It was also less likely to be affected by unimportant but specific aspects of item content. CFA using parceling of the items in both the 75 item and Rasch-produced 44 item versions with no differential item functioning resulted in strong fit statistics, indicating that both versions of the scale are best viewed as a unidimensional measure of systemizing. It should also be remembered that these results were independent of method of parceling, as both random and the single factor method (Aluja & Blanch, 2004) of parceling produced acceptable fit statistics.

As Nettle (2007) highlights systemizing appears to emerge as a cognitive style, whereby people are drawn to understanding causal relationships in non-social domains, which can explain the greater male interest and performance in subjects such as science and mathematics (Geary, Saults, Liu, & Hoard, 2000), and the extreme end of the typical male profile in ASC (Baron-Cohen, 2002). No data are available that assess whether an intact or increased drive to systemize holds true right across the whole autism spectrum. However, repetitive behaviors (including what is described as 'stimming' (Wing, 1997) as well as collecting obsessions, lining things up, and being drawn to repeating patterns) may be the hallmark of strong systemizing in individuals with ASC and severe learning disability (Baron-Cohen, 2009).

The current research using two different but complementary perspectives and statistical methods (Rasch modeling and CFA), both highlight that the construct of systemizing is unidimensional and can be measured using the SQ-R. Thus, using a total score on the SQ-R adequately represents and quantifies an individual's drive to systemize. These data support Baron-Cohen's (2002) construct of systemizing as a drive. There are items in the SQ-R that enquire about preferences, understanding, and to some extent abilities, but these

https://doi.org/10.1017/sjp.2015.19 Published online by Cambridge University Press

all appear to relate to one single underlying construct of a drive to construct systems, to predict the behavior of a system, and to control it. This study does not address whether a relationship exists between systemizing and intelligence. Previous research suggests this is not the case (with intelligence measured by the Baddeley 3 minute reasoning test) (Ling et al., 2009). Further research is required with more standardized measures of IQ, to properly address this question. Systemizing (in conjunction with empathy measures) is a better predictor than sex of whether a student opts to study science vs. humanities (Billington et al., 2008), suggesting that cognitive style (independent of sex) underlies choice of degree. Whether the drive to systemize is directly related to an *ability* to systemize, remains untested. Do people with a strong drive towards systemizing always demonstrate high levels of performance on measures that are designed to directly measure and quantify individual differences in systemizing? Can these individuals be separated from individuals with a high drive to systemize, but low levels of performance on systemizing measures? Robust performance measures of systemizing are required to address these questions. Further, performance measures of systemizing that are applicable to the whole population, including people with autism and intellectual disability should be developed. In summary, two different statistical approaches support the idea that the SQ is an appropriate measure of the construct of systemizing which can be measured along a single dimension.

References

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

Aluja A., & Blanch A. (2004). Replicability of first-order 16PF-5 factors: An analysis of three parceling methods. *Personality and Individual Differences*, *37*, 667–677. http://dx.doi.org/10.1016/j.paid.2003.10.001

Andersen E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, *42*, 69–81. http://dx.doi.org/10.1007/BF02293746

Arbuckle J. L. (2006). Amos (Version 7.0). Chicago, IL: SPSS.

Auyeung B., Allison C., Wheelwright S., & Baron-Cohen S. (2012). Brief report: Development of the adolescent empathy and systemizing quotients. *Journal of Autism and Developmental Disorders*, *42*, 2225–2235. http://dx.doi.org/10.1007/s10803-012-1454-7

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*, 1509–1521. http://dx.doi.org/10.1007/s10803-009-0772-x

Baron-Cohen S. (2002). The extreme male brain theory of autism. *Trends in Cognitive Science*, *6*, 248–254. http://dx.doi.org/10.1016/S1364-6613(02)01904-6

Baron-Cohen S. (2009). Autism and the Empathizing–Systemizing (E-S) theory. In P. D. Zelazo, M. Chandler, & E. Crone (Eds.), *Developmental social cognitive neuroscience* (pp. 125–138). New York, NY: Psychology Press.

Baron-Cohen S., Richler J., Bisarya D., Gurunathan N., & Wheelwright S. (2003). The systemizing quotient: An investigation of adults with Asperger syndrome or high-functioning autism, and normal sex differences. *Philosophical Transactions of the Royal Society London B Biological Sciences*, *358*, 361–374. http://dx.doi.org/10.1098/rstb.2002.1206

Baron-Cohen S., Wheelwright S., Griffin R., Lawson J., & Hill J. (2002). The exact mind: Empathising and systemising in autism spectrum conditions. In U. Goswami (Ed.), *Handbook of Cognitive Development*. Oxford, UK: Blackwell.

Billington J., Baron-Cohen S., & Bor D. (2008). Systemizing influences attentional processes during the Navon task: An fMRI study. *Neuropsychologia*, *46*, 511–520. http://dx.doi.org/10.1016/j.neuropsychologia.2007.09.003

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

Comrey A. L. (1988). Factor analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology*, *56*, 754–761. http://dx.doi.org/10.1037//0022-006X.56.5.754

Cook K. F., Kallen M. A., & Amtmann D. (2009). Having a fit: Impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. *Quality of Life Research*, *18*, 447–460. http://dx.doi.org/10.1007/s11136-009-9464-4

Floyd F. J., & Widaman K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, *7*, 286–299. http://dx.doi.org/10.1037//1040-3590.7.3.286

Geary D. C., Saults S. J., Liu F., & Hoard M. K. (2000). Sex differences in spatial cognition, computational fluency, and arithmetical reasoning. *Journal of Experimental Child Psychology*, *77*, 337–353. http://dx.doi.org/10.1006/jecp.2000.2594

Hall R. J., Snell A. F., & Foust M. S. (1999). Item parceling strategies in SEM: Investigating the subtle effects of unmodeled secondary constructs. *Organizational Research Methods*, *2*, 757–765. http://dx.doi.org/10.1177/109442819923002

Kenny D. A., & McCoach D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*, 333–351. http://dx.doi.org/10.1207/S15328007SEM1003_1

Landis R. S., Beal D. J., & Tesulk P. E. (2000). A comparison of approaches to forming composite measures in structural equation models. *Organizational Research Methods*, *3*, 186–207. http://dx.doi.org/10.1177/109442810032003

Linacre J. M. (1998). Structure in Rasch residuals: x (PCA)? *Rasch Measurement Transactions*, *12*, 636.

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

Ling J., Burton T. C., Salt J. L., & Muncer S. J. (2009). Psychometric analysis of the systemizing quotient (SQ) scale. *British Journal of Psychology*, *100*, 539–552. http://dx.doi.org/10.1348/000712608X368261

Mavranezouli I., Brazier J. E., Young T. A., & Barkham M. (2011). Using Rasch analysis to form plausible health states amenable to valuation: The development of CORE-6D from a measure of common mental health problems (CORE-OM). *Quality of Life Research*, *20*, 321–333. http://dx.doi.org/10.1007/s11136-010-9768-4

Marsh H. W., Hau K., Balla J. R., & Grayson D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factory analysis. *Multivariate Behavioral Research*, *33*, 181–220. http://dx.doi.org/10.1207/s15327906mbr3302_1

Nettle D. (2007). Empathizing and systemizing: What are they, and what do they contribute to our understanding of psychological sex differences? *British Journal of Psychology*, *98*, 237–255. http://dx.doi.org/10.1348/000712606X117612

Nunnally J. C., & Bernstein I. H. (1994). *Psychometric theory*. New York, NY: McGraw Hill.

Perline R., Wright B. D., & Wainer H. (1979). The rasch model as additive conjoint measurement. *Applied Psychological Measurement*, *3*, 237–255. http://dx.doi.org/10.1177/014662167900300213

Rasch G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: The University of Chicago Press.

Scheuneman J. D., & Subhiyah R. G. (1998). Evidence for the validity of a Rasch model technique for identifying differential item functioning. *Journal of Outcome Measurement*, *2*, 33–42.

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

Tristan A. (2006). An adjustment for sample size in differential item functioning analysis. *Rasch Measurement Transactions*, *20*, 2.

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*, 1823–1832. http://dx.doi.org/10.1007/s10803-006-0316-6

Wakabayashi A., Baron-Cohen S., Wheelwright S., Goldenfeld N., Delaney J., Fine D., … Weil L. (2006). Development of short forms of the Empathy Quotient (EQ-Short) and the Systemizing Quotient (SQShort). *Personality and Individual Differences*, *41*, 929–940.

Wheelwright S., Baron-Cohen S., Goldenfeld N., Delaney J., Fine D., Smith R., … Wakabayashi A. (2006). Predicting autism Spectrum Quotient (AQ) from the Systemizing Quotient-Revised (SQ-R) and Empathy Quotient (EQ). *Brain Research*, *1079*, 47–56. http://dx.doi.org/10.1016/j.brainres.2006.01.012

Wing L. (1997). *The autistic spectrum*. London, UK: Pergamon.

Wright B. D., & Stone M. H. (2004). *Making measures*. Chicago, IL: The Phaneron Press.

Yang C., Nay S., & Hoyle R. H. (2010). Three approaches to using lengthy ordinal scales in structural equation models: Parceling, latent scoring, and shortening scales. *Applied Psychological Measurement*, *34*, 122–142.

https://doi.org/10.1017/sjp.2015.19 Published online by Cambridge University Press