Published as: Murray, A. L., Booth, T., McKenzie, K., & Kuenssberg, R. (2015). What Range of Trait Levels Can the Autism-Spectrum Quotient (AQ) Measure Reliably? An Item Response Theory Analysis. *Psychological Assessment,* Online First.

What range of trait levels can the Autism-Spectrum Quotient (AQ) measure reliably? An item response theory analysis

Abstract

It has previously been noted that inventories measuring traits that originated in a psychopathological paradigm can often reliably measure only a very narrow range of trait levels that are near and above clinical cut-offs. Much recent work has, however, suggested that autism spectrum disorder (ASD) traits are on a continuum of severity that extends well into the non-clinical range. This implies a need for inventories that can capture individual differences in autistic traits from very high levels all the way to the opposite end of the continuum. The Autism-Spectrum Quotient (AQ) was developed based on a closely related rationale but there has to date been no direct test of the range of trait levels that the AQ can reliably measure. To assess this, we fit a bi-factor item response theory model to the AQ. Results suggested that AQ measures moderately low to moderately high levels of a general autistic trait with good measurement precision. The reliable range of measurement was significantly improved by scoring the instrument using its four-point response scale, rather than dichotomising responses. These results support the use of the AQ in non-clinical samples, but suggest that items measuring very low and very high levels of autistic traits would be beneficial additions to the inventory.

Published as: Murray, A. L., Booth, T., McKenzie, K., & Kuenssberg, R. (2015). What Range of Trait Levels Can the Autism-Spectrum Quotient (AQ) Measure Reliably? An Item Response Theory Analysis. *Psychological Assessment,* Online First.

Introduction

Although concepts surrounding autism spectrum disorder (ASD) traits originated within a psychopathological paradigm, there is increasing consideration and acceptance of the idea that ASD traits may exist on a continuum that spans both clinical and non-clinical levels (e.g. Frazier et al., 2010; Lundström et al., 2012). In this view, individuals who receive a clinical diagnosis of ASD may simply be at the extreme end of this continuum, rather than manifesting some qualitatively distinct condition (Austin, 2005).

A corollary of this view is that there is meaningful variation in autistic traits to be measured and understood below a clinical threshold for ASD. This being the case, it is important for empirical studies to capture this sub-clinical variation in addition to variation at clinical levels. It has been noted, for example, that the statistical power to detect associations with putative genetic or environmental causes can be improved by utilising samples which include respondents exhibiting variation at both clinical and sub-clinical levels (Lundström et al., 2012). Conversely, failing to measure the full range of an autistic trait continuum could lead to an under-statement of these and other associations due to range restriction (Murray, McKenzie, Kuenssberg & O'Donnell, 2014).

It is the goal of capturing a fuller range of autistic trait levels than is traditionally assessed by purely clinical measures that has motivated the development of assessments of 'broader autism phenotype' or 'autistic-like traits' (e.g. Wheelright, Auyeung, Allison, & Baron-Cohen, 2010). Assessments developed with this goal include the Autism Spectrum Quotient (AQ; Baron-Cohen , Wheelright. Skinner, Martin & Clubley, 2001), the Social Responsiveness Scale (SRS; Constantino & Gruber, 2005), and the Broader Autism Phenotype Questionnaire (BAPQ; Hurley, Losh, Parlier, Reznick, & Piven, 2007). Consistent

with the idea of autistic traits as a dimension, these scales focus on measuring the extent or severity of autistic traits, rather than on determining category membership in an ASD versus no ASD group.

Concerns have, however, been raised separately about the ability of inventories measuring traits originating in a psychopathological paradigm to effectively capture the full range of these constructs. Several authors have noted that inventories measuring clinical traits often contain items covering only a limited range of trait levels; specifically, those in the clinical range (e.g. Meijer & Egberink, 2012; Reise & Waller, 2009; Schwabe & van den Berg, 2014). This limited range becomes apparent when the inventories are analysed using item response theory models. In these models items tend to show high discrimination parameters but difficulty parameters clustered in the clinical range. Discrimination parameters are indices of how well items can differentiate between individuals of different trait levels and difficulty parameters are indices of where on the trait continuum items can do this with the greatest degree of precision. The particular combination of high discrimination and closely clustered difficulty parameters results in so-called 'peaked' tests whereby levels close to some clinical cut-off are measured with very high measurement precision but non-clinical levels are measured with very limited precision (Kang & Waller, 2005; Reise & Waller, 2009). Inventories exhibiting these kinds of properties, therefore, provide reliable measures of a trait for a very limited range of levels and may make them ideal for classifying or diagnosing individuals as affected by the clinical trait of interest, but not for measuring sub-clinical levels of that trait (Thomas, 2011).

These kinds of scaling properties are associated with clinical traits because of the particular manner in which items are commonly selected. Often, items are chosen with the goal of maximising test discrimination which has traditionally meant selecting those that

show strong inter-correlations. Where this becomes problematic, is in the fact that item inter-correlations are attenuated when they differ in their response distributions or difficulty (Embretson & Reise, 2000). As a result, selecting items to have high discrimination can result in a restriction in the range of their difficulty parameters with attendant consequences for the range of trait levels that can be reliably measured. Similarly, there is a tendency to select items which refer to the most severe manifestations of the trait because it is these that can best differentiate those with and without the trait (van den Oord, Pickles & Waldman, 2003). These kinds of items are also likely to be the least ambiguous markers of the trait. For example, whereas severe social deficits have strong face validity in relation to autistic traits, more subtle markers such as preferring to spend time at a library than a party are less obviously related to ASD. In part, this may be a function of the fact that the conceptualisation of sub-clinical psychopathological traits is a relatively recent development and test developers may, therefore, have much less previous theory and evidence to draw on when writing items. They may instead rely on accumulated knowledge from the much longer history of the study of clinical-level psychopathological traits. In sum, for various reasons traditional test development and evaluation procedures in the clinical domain may implicitly restrict the range of clinical traits that are ultimately reliably captured by psychometric inventories.

Another consequence of this restricted range of reliable measurement is the possibility of spurious statistical results when inventory scores are employed to test substantive hypotheses. When scores affected in this way are used as outcomes in a moderation analysis they can lead to spurious detection of interactions or the masking of genuine moderation effects. It occurs because the compression of systematic variance at one end of the trait continuum due to unreliable measurement is not completely statistically distinguishable from

an interaction effect. This phenomenon has been demonstrated in phenotypic interactions in moderated multiple regression (Kang & Waller, 2005), factorial ANOVA (Embretson, 1996), and gene-environment interactions in behaviour genetic models (e.g. Schwabe & van den Berg, 2014) where it presents an important methodological and interpretative challenge. If the range of reliable measurement is restricted severely enough to the point where serious floor or ceiling effects occur, it can lead to the detection of a GxE effect in the opposite direction to the true effect (Murray, Molenaar, Johnson & Krueger, submitted).This issue is especially pertinent given the recent interest in GxE interactions as important etiological mechanisms in psychopathological phenotypes (e.g. Caspi & Moffit, 2006).

Given these concerns and the growing interest in measuring the full breadth of a hypothesised continuum of autistic traits, it was the aim of the current study to use an item response theory approach to assess the extent to which a popular measure of autistic traits provides good coverage of the hypothesised continuous ASD phenotype. We focus on the Autism Spectrum Quotient (AQ; Baron-Cohen et al., 2001) because it is one inventory that was created with the goal of placing an individual on a continuum from normality to autism and, therefore, acknowledges the need to develop new inventories for this purpose, rather than simply administering existing clinical measures. The inventory was developed for use in adults of normal intellectual ability in the general population. It is not intended to be a diagnostic instrument; rather, its purpose is to quantify levels of autistic traits. It comprises 50 items organised into 5 domains labelled 'Social Skills', 'Attention Switching', 'Attention to Detail', and 'Imagination' and thus measures, in addition to the areas of the classical triad, a number of features that the literature suggests are associated with ASD. The uptake of the AQ in empirical research has been extensive with the original validation paper having been

cited more than 1500 times at time of writing. Overall, it is very widely used in both research and clinical practice as a method of quantifying autistic traits (Ruzich et al., 2015).

Some preliminary evidence has suggested that the AQ successfully captures systematic variation in both the clinical and non-clinical ranges of ASD traits. Summing the 50 items of the AQ yields a near-normal distribution of observed scores in the general population (Baron-Cohen et al., 2001), and substantive variance (i.e. non-error variance) in the measure is greater when individuals from a clinical and non-clinical sample are combined relative to measuring the AQ in either sample alone (Murray, McKenzie, Kuenssberg & O'Donnell, 2014). No study has, however, directly addressed range of autistic trait levels that can be reliably measured by the AQ. This was, therefore, our aim in the current study.

Method

Participants

Data came from three archival samples: one clinically diagnosed sample and two control samples. These were combined for the current study in order to ensure a broad range of trait levels were represented. This gave us a total sample size of 579 (208 males, 370 females, 1 other gender). A small amount of data were missing (<1%), which was dealt with using maximum likelihood estimation. Ethical approval was obtained from the relevant ethics committee in all cases.

Invariance sample

One hundred and sixty seven participants came from a previous psychometric study of the AQ (Murray, Booth, McKenzie, Kuenssberg & O'Donnell. 2014). Participants were recruited from the university community and online. They comprised 40 males and 127

females with a mean age of 30.0 (SD=11.3). This sample has also been previously used in the real data example in the simulation study by Murray, McKenzie et al. (2014) and in Murray, Kuenssberg, McKenzie, Booth (2015).

Emotion recognition sample

Ninety eight participants came from an ongoing study of emotion recognition and ASD traits and were also used in the real data example in Murray, McKenzie et al. (2014) and in Murray, Kuenssberg et al. (2015). These participants were recruited online and from the university community and included 27 males, 70 females and 1 participant who described their gender as 'other'. The mean age of the sample was 31.0 (SD=12.5). None reported having a formal diagnosis of ASD.

Sex differences sample

One hundred and sixty six participants came from an ongoing study of sex differences in ASD traits in individuals without a clinical diagnosis of ASD. This sample was also used in the real data example in Murray, McKenzie et al. (2014) and in Murray, Kuenssberg et al. (2015). Participants were recruited online and were composed of 132 females and 34 males. The mean age of the sample was 27.1 (SD=12.3). None reported having a formal diagnosis of ASD.

Clinical sample

One hundred and forty eight participants came from a sample of individuals with a clinical diagnosis of ASD. Individuals were diagnosed with either Asperger's syndrome or high functioning autism. The sample has been utilised in several previous studies (Booth, Murray et al., 2013; Murray, Booth et al., 2014; Kuenssberg , Murray, Booth & McKenzie,

2014; Murray, McKenzie et al., 2014; Murray, Kuenssberg et al., 2015) and is comprehensively described in terms of data collection procedures and sample composition in Kuenssberg et al. (2014). Data were obtained from case notes from ASD and psychological services. Data include 107 males and 41 females with a mean age of 33.3 (SD=10.7).

Measures

The autism spectrum quotient is a 50 item questionnaire developed to measure ASD traits in adults of normal intellectual ability (Baron-Cohen et al., 2001). The items are organised into 10 item scales measuring the traits of Social Skills, Attention Switching, Attention to Detail, Communication, and Imagination. These content areas were influenced by both the diagnostic criteria for ASD at the time of its development as well as commonly associated features of ASD. At the time of its development, the classical triad of ASD formed the basis of diagnosis which comprises social interaction impairment, communication difficulties, and restricted repetitive behaviours (APA, 1994). Now, following the publication of DSM 5, the diagnostic criteria for ASD have been updated (APA, 2014). A major difference is that in DSM 5, the social interaction and communication domains were combined into a single 'Social Communication' domain. This change was based on empirical evidence that the Social and Communication domains tended to show strong correlations with one another. The relevance of these changes for the use of the AQ has not, to our knowledge, been formally examined. The AQ is, however, not intended to be used in the context of diagnosis; only as a means of quantifying autistic traits, therefore, the changes in diagnostic criteria are not likely to significantly impact its use or interpretation.

Items of the AQ are phrased in terms of behavioural tendencies and preferences, and the respondents are asked to answer items using a four point scale with options 'strongly

agree', 'slightly agree', 'slightly disagree' and 'strongly disagree', however, the majority of previous studies have collapsed this scale into a dichotomous '0' vs '1' scoring scheme. In order to evaluate the AQ as it is used in practice, we also follow this dichotomising procedure. This practice is not ideal because it discards information about trait values. To gauge the impact of dichotomising we repeated analyses scoring items using the full four point scale and compared these to the results using the dichotomised scoring scheme. Half of the items are coded in the forward direction and half are coded in the reverse directionn. For the current study, all items were coded such that higher scores represented being higher on an autistic trait. Abbreviated item contents are provided in Table 1. Exact item wordings can be found in the appendix of Baron-Cohen et al. (2001).

The psychometric properties and correlates of the AQ, which have been examined across a large number of studies, are generally supportive of its use as a measure of autistic traits. Consistent with what would be expected of a measure of autistic traits, males tend to score higher than females; science students tend to score higher than non-science students; and individuals with a clinical diagnosis of ASD tend to score higher than controls with no clinical diagnosis of ASD (Austin, 2005; Baron-Cohen et al. 2001; Hoekstra, Bartels, Cath, Boomsma, 2008). Scores also correlate with autistic features such as empathising (Wheelright et al., 2006), emotion recognition (Baron-Cohen, Wheelright, Hill, Raste & Plumb, 2001) and alexithymia (Liss, Mailloux & Erchull, 2008). With regards to its divergent validity, the AQ appears to capture traits that are distinct from those covered by major contemporary models of personality (Austin, 2005; Wakabayashi, Baron-Cohen & Wheelright, 2006). It also appears to be sensitive to the features of ASD specifically, rather than those of other clinical disorders such as schizophrenia (Naito, Matsui, Maeda & Tanaka, 2010; Spek & Wouters, 2010) and attention deficit hyperactivity disorders (ADHD; Sizoo et

al., 2009) albeit not with levels of discriminability that would support reliance on it for differential diagnosis in clinical practice.

Statistical Procedure

There has been some disagreement as to the most appropriate dimensionality of the AQ in terms of the number and content of the specific factors (e.g. Lau, Kelley & Peterson, 2013). Nonetheless, previous research supports the broad contention that both a general factor and more specific symptoms and features of ASD are captured by the items of the AQ (Hoekstra et al., 2011; Murray et al., 2015). This is also reflected in the design of the instrument which can be scored as a single dimension or as 5 subscales reflecting specific features of ASD.

In practice the AQ is most commonly scored as the sum of its 50 items and used as a measure of a general autistic trait; subscale scores are used less frequently. It is, therefore, on the reliability of measurement of scores for this general trait that we primarily focus. Another reason to focus on the general trait is that a previous study in the same sample identified that items of the AQ mostly reflect a general trait, rather than the specific traits measured by each subscale (Murray et al., 2015). Thus, a large part of the reliability of the subscales scores is in fact due to a general trait. After partialling out the variance due to the general trait, the subscales showed weak unique reliability with omega hierarchical values between only .05 for the Communication and Social Skills subscales and .67 for the Attention to Detail subscales.

It is, nonetheless important to acknowledge and model the influence of the clustering of items within symptom-specific subscales. Such clustering creates violations of local independence which inflates estimates of the reliability of items and tests (e.g. Chen &

Thissen, 1997). We, therefore, used a bi-factor measurement model with a normal ogive link function to account for this clustering (Gibbons et al., 2007). Here, items were modelled as a function of both a general factor and a specific factor, with all of these factors orthogonal to one another. The bi-factor structure utilised in the current study specified all items as influenced by a general autistic factor as well as a specific factor corresponding to the AQ sub-scale from which it came. In this specification one threshold was estimated for each item, representing the combination of general and specific trait level necessary to endorse that item. In addition, each item had a slope for the general factor and the relevant specific factor. These general factor slopes were conditional on the specific factors and vice versa, allowing the conditional item information (and thus measurement precision) in the direction of the general factor to be computed. In doing so, the bi-factor structure addressed the problem of violations of local independence due the presence of specific factors.

A key advantage of parametric IRT is that it acknowledges that the precision of measurement for a trait is not equal across all levels that trait. For dichotomously scored items, information is greatest at the difficulty parameter and decreases above and below this. When an item measures only one trait, the item information function is given by:

$$I(\theta) = \frac{[P'(\theta)]^2}{P(\theta)[1 - P(\theta)]}, \quad (1)$$

where $P(\theta)$ is the item response function and $P'(\theta)$ is its first derivative. However, in models such as the bi-factor model in which items measure more than one trait, a multi-dimensional generalisation of the item information function must be considered (Reckase, 1991). To obtain the item information function for a specific trait (e.g. the general factor in a bi-factor

model) the direction of information must be taken into account. Conditional item information in that direction can then be described by:

$$I_d(\theta) = \frac{[\nabla_d P(\theta)]^2}{P(\theta)[1 - P(\theta)]} \quad (2)$$

where $\nabla_d P(\theta)$ is the gradient in direction d**.** When information in the direction of the general factor is computed, it is conditional on the specific factors and, therefore, addresses the problem of local dependencies due to the presence of the specific factors. Then local independence can be assumed and the item information functions summed across all items in the test to get the test information function for the general factor. Conditional item and test information are inversely related to the standard error of measurement, specifically, the conditional standard error is the reciprocal of the square root of the conditional information.

Thus, computing the test information function allows an evaluation of the locations along the latent trait continuum that are relatively more or less precisely measured by the test as a whole. Highly peaked curves suggest that the test measures reliably within a narrow range of the trait. This property is undesirable for a test such as the AQ which purports to measure autistic traits into the non-clinical range. Conversely, curves that depict high information values across a broader range of the continuum are more useful for measuring a broader range of the trait. As an approximate gauge of the breadth of reliable measurement, we inspected the points at which the information falls below 10 (corresponding to a classical test theory reliability of 0.90) can suggest how wide the range of reliable measurement is. All models were estimated in *Mplus 6.11* using maximum likelihood estimation (Muthén & Muthén, 1998-2013).

Results

Dichotomous scoring

On initially fitting a bi-factor IRT model to the dichotomously scored AQ items, the model failed to converge. The problem appeared to involve item 38 which had a large general factor slope and could not be simultaneously loaded on the Communication subscale without estimation problems. We, therefore, specified this item to load only on the general factor. Parameter estimates for this model are provided in Table 1.

The slope parameters indicate how well the item differentiates between individuals of different levels of the general and specific factors. Based on their general factor slopes, items relating to social situations were most discriminating for the general factor. For example, the largest general factor slope parameters of 3.29 and 2.18 for items 44 and 11 respectively were both from items in the Social Skills scale. The location parameters indicate how high an individual needs to be on their combination of (general and specific) autistic traits to endorse that item. For example, results suggest that item 12 referring to noticing small noises (location= -1.11) does not require high levels of autistic traits to be endorsed. On the other hand, item 44 referring to (a lack of) enjoyment of social occasions (location= 1.02) requires relatively high levels of autistic traits to be endorsed. Overall, the distribution of location parameters suggested that the items of the AQ measured a relatively wide range of autistic trait levels.

The test information curve for the general factor is provided in Figure 1. The maximum information was 36.0 at a latent trait value of 0.23. The lowest latent trait value for

which the information value was above 10 (corresponding to a classical test theory reliability of 0.90) was -1.18 and the highest 1.52. Therefore, the AQ, when scored dichotomously, is best at capturing moderately low to moderately high levels of a general autistic trait while measuring very low and very high autistic traits with limited precision.

We also estimated the conditional test information curves for each of the specific factors; however, conditional test information did not exceed 10 at any trait value for any specific factor. This suggests that these should not be used as measures of specific ASD factors.

Four-point response scale scoring

The model for the four point response scale converged to reasonable parameter values without the need to make any model modifications. Parameter estimates are provided in Table 2. The pattern of slope and location parameters is similar to that observed when the items are dichotomously scored.

The test information curve for the general factor with items scored on a four-point scale is provided in Figure 2. The maximum information was 44.32 at a latent trait value of 0.53. The lowest latent trait value for which the information value was above 10 was -2.45 and the highest was at 2.55. The reliable range of measurement of the AQ was, therefore, substantially extended in both directions by scoring the items on a four point scale as compared to a dichotomous scale.

On estimating the conditional test information functions for the specific factors, there were again no latent trait values for which the information exceeded 10. Thus, even scoring

the items as a four point scale did not overcome the low reliability of the subscales as measures of specific factors.

Discussion

In the current study, we used a bi-factor IRT model to assess the extent to which the AQ was subject to a limitation commonly observed in the measurement of traits originating in a psychopathological paradigm: that of measurement precision localised to the clinical range of the trait. Analyses suggested that the AQ best captures autistic traits within the moderately low to moderately high range. However, it could measure very high and very low levels of autistic traits with limited measurement precision. This suggests that, unlike many measures developed in, or influenced by, a psychopathological paradigm (Reise & Waller, 2009), the AQ is best suited to capturing individual differences in autistic traits in non-clinical populations. Of note, the range of latent trait values that could be reliably measured by the AQ was much greater when items were scored on a four point response format as compared to more commonly used dichotomous response format.

The fact that the AQ captures a general autistic trait best at moderate levels is in keeping with its design as an instrument for placing an individual on a continuum from autism to normality. It is also consistent with the content of items which refer primarily to relatively mild manifestations of autistic traits. For example, when referring to social difficulties, respondents tend to be asked to agree with a statement that a situation or task is 'hard' or 'difficult' rather than, for example, 'impossible', 'extremely uncomfortable', or 'extremely upsetting.' In addition to the mild wording of items, severe autistic traits are not represented in the test. For example, restrictive repetitive behaviours characteristic of clinical

levels of autism are not referred to. Thus, the range of reliable measurement of the AQ could likely be extended with the addition of items referring to more severe autistic traits.

A similar argument can be made for extending the reliable range of measurement to lower levels of autistic traits. For example, the wording of most items suggests that failing to endorse them would not require a person to have characteristics strongly opposed to autistic traits. Therefore, including items which reflect not only a lack of autistic traits, but their opposite may help to extend the range of reliable measurement towards the lower end of the trait continuum.

In terms of maximising the reliable range of measurement using the existing set of items, our results suggested a major benefit of scoring the items using the four-point response scale provided for respondents. Typically, in empirical applications, the items are dichotomised prior to summing; however, this practice restricts the reliable range of measurement of the scale substantially. At low trait levels, using the 4 point response scale allows traits down to -2.45 SDs to be reliably measured as compared to only -1.18 SDs when collapsing responses to a 2 point scale. At high trait levels, using the 4 point response scale allowed reliable measurement of trait levels up to 2.55 SDs compared with only 1.52 using a dichotomous scoring scheme. Thus, on the basis of our results, we would strongly recommend using the 4 point scale over the 2-point scale.

It would also be beneficial to examine whether the same kinds of gains can be achieved in the AQ-10 by using a 4-point response format. The AQ-10 is a 10 item subset of the AQ recommended as a brief screen for ASD in cases where ASD is already suspected (Allison, Auyeung & Baron-Cohen, 2012). The AQ-10 also uses a dichotomous scoring scheme where the original four-point response format is collapsed into a 2-point scale for the

purposes of scoring. Then, individuals scoring above a cut-off of 6 are classified as showing levels of ASD traits that would merit referral for formal assessment. Given its function as a screening tool and the goal that it be used by frontline professionals, the ease of a simple dichotomous scoring format is an advantage. Further, in contrast to the AQ and also consistent with its function as a screening tool, the most important consideration is its ability to discriminate between cases and non-cases rather than its range of reliable measurement. It is nonetheless likely that using the four-point scale could outweigh the usability benefits of a dichotomous scoring scheme as it is likely that using the four-point scale for scoring could improve the sensitivity and specificity of the instrument.

The results concerning the difference between the two- and four-point response format also raise the question of whether offering respondents a fifth response category would further improve the reliable range of measurement of the AQ itself. Given that it currently has a four-point scale, this would most sensibly entail including a middle option such as 'neither agree nor disagree'. Previous research would suggest that including such an option could have mixed results. While not offering a middle option can promote missing responses (whenever, for example, a respondent is unsure about their answer), respondents do not always use middle options as indicators of intermediate trait levels (Hernández , Drasgow & González-Romá, 2004). Rather, selecting the middle response option can reflect reticence, uncertainty, confusion, that the question is not applicable, that the answer to the question is dependent on circumstances or context and a whole host of other unintended factors (e.g. Kulas & Stachowski, 2009, 2013; Kulas, Stachowski & Haynes, 2008). This can lead to increased measurement error and systematic errors in estimates of trait levels if choosing the middle response option is related to levels of autistic traits as has been observed in relation to some other personality traits (Murray, Booth & Molenaar, 2015). There is also the related

question of the optimal number of response options. Previous research has suggested that there is no single optimal number of response options that applies across all inventories; however, a general observation is that reliability tends to increase with number of response options up to a point and then decreases again (e.g. Lozano, García-Cueto & Muñiz, 2008). In addition, it is not only the number of response options that affects reliability, but the verbal labels given to them (e.g. 'agree') (Rammstedt & Krebs, 2007). Further research will be required to establish what the optimal number and labelling of response options is for the AQ.

Results also highlighted the potential difficulty of reliably measuring specific ASD factors. Controlling for the general factor, the slopes for the specific factors were relatively small and on estimating the test information functions for the corresponding subscale scores, it was found that there was no range of specific trait values for which item information was above 10. Indeed, the pattern of slope parameters suggested that when the sub-scale is administered, primarily it is a general ASD trait that is being measured (also see Murray et al., 2015). Given that the AQ was developed and primarily administered to measure a general level of ASD, rather than being focussed on specific facets, this is in line with the design and use of the instrument.

Many items also had small slope parameters for the general factor, especially those from the 'Attention to Detail' domain. One possibility is that these items are contextualised to kinds of information that are overly specific (e.g. interest in dates versus phone numbers) with the result that idiosyncrasies of interests play too large a role in item responding. For example, with regards to obsessive interests in ASD, it has been noted that some individuals may develop a particular interest in dates, while others may become more interested in phone numbers, prime numbers, or other categories of highly specific information (e.g. Klin,

Danovitch, Merz, & Volkmar, 2007). Another possibility is that the behaviours to which they refer are simply less relevant to ASD than the other items included in the AQ.

Another factor likely contributing to the small general factor slopes for some items is the diversity of behaviours to which the items refer. This was likely achieved by developing items based, not only on the triad that formed the basis of clinical diagnosis at the time, but also on established associated features (Baron-Cohen et al. 2001). Diversity of item content helps to ensure that a scale is not limited to measuring a specific, overly narrow construct (Meijer & Egberink, 2012). As noted in the introduction, a danger in developing scales to measure clinical constructs is selecting items to have high discrimination values, leading to a scale with many items that are very similar, perhaps even essentially repetitions of one another. This kind of item selection procedure can have an adverse effect both on the content validity of the scale (it measures too narrow a construct) and on its reliable range of measurement (it measures only a narrow range of trait levels with sufficient reliability).

There are, however, still some limitations of the AQ. Our analyses suggested that even using a 4-point response format, there were limits to its reliable range of measurement. Furthermore, it can be used only in adults with normal intellectual ability, which excludes a large number of individuals with ASD who have co-morbid intellectual disability. The extent to which other measures of autistic traits have a good range of reliable measurement has, to our knowledge, not been assessed. It will, therefore, be a potentially important future direction to evaluate, using similar methodologies to that presented in the current study, which measures of autistic traits are capable of reliably measuring the largest range of autistic traits. This would help inform choices of measures for studies aiming to assess autistic traits across a broad range of autistic trait levels such as family or other large epidemiological studies.

Limitations

The ability to assess the relative contribution of general and specific factors to item responding is an advantage of the bi-factor model (e.g. Reise, Morizot & Hays, 2007). Another advantage is that any local dependence due to the fact that items are influenced by both general and specific factors can be accounted for. When such multi-dimensionality of responding occurs, fitting either a single uni-dimensional model to all 50 items, or fitting individual uni-dimensional models to each of the sub-scales can lead to distorted parameter estimates. However, it is also possible that the bi-factor model utilised in the current study was mis-specified in some important respect. For example, in terms of the specific factors measured by the AQ, there exists no clear agreement on the appropriate factorial structure of the AQ and its derivatives, with different studies suggesting markedly different numbers and contents of factors (e.g. see Lau et al., 2013). This is not an issue specific to the AQ, but one characteristic of factor analytic research in which there is huge variability in the factor structures presented by different authors analysing the same inventory. However, in specifying the structure corresponding with the instruments scoring scheme, we were able to test the AQ in a manner consistent with how it was designed and is used in practice.

In terms of the limitations of the current study, the sample was not a random draw from the relevant population and was, therefore, not representative of the general population in terms of characteristics such as ASD prevalence, gender ratios, or age. With regards to gender ratios, while males are more likely to exhibit trait levels in the clinical range and females in the non-clinical range, our non-clinical samples still included a disproportionate number of females while the clinical sample included a disproportionate number of males.

However, the study did have the advantage of including individuals both with and without a clinical diagnosis of ASD, giving a good range of ASD trait levels, which was of key importance for the particular research question we were focussed on. Nonetheless, future studies should aim to replicate these results in larger, more population representative samples.

Another limitation is our inability to account for possible systematic relations between the validity of self-report and autistic traits. There is, for example, a possibility that individuals with high levels of autistic traits have poorer self-insight and, thus, provide unreliable reports of their behaviour and preferences when responding to the AQ items (Baron-Cohen et al., 2001; Johnson, Filliter & Murphy, 2009). While the current methodology can detect increased measurement error at high levels of autistic traits, it cannot identify whether this derives from the item contents being less relevant at such levels or individuals providing less reliable self-reports.

Finally, given sex differences in prevalence and presentation in ASD, it is a potential limitation that we did not assess differential item functioning (DIF) by sex (Lai, Lombardo, Auyeung, Chakrabarti & Baron-Cohen, 2015). It is, in principle, possible that DIF by sex could lead to a reliable range of measurement that differed across males and females. One previous study has examined DIF by sex in a subset of the AQ; namely, the 10 items of the AQ that comprise the AQ-10 screening tool. It found that while individual items showed bias in one or other direction, these cancelled out at the level of the test thus giving no overall differential test functioning (Murray, Allison, Auyeung, Smith, Baron-Cohen, Booth, 2015). It is not known whether this kind of pattern would generalise to the entire AQ, making it a potentially important future direction to determine whether the reliable range of measurement for the AQ differs for males and females.

Conclusions

Consistent with its design and in contrast to many other scales measuring traits originating in the psychopathological paradigm, the AQ provides a reliable measure of autistic traits for moderately low to moderately high levels of a general autistic trait. This supports its utility in non-clinical samples. In addition, scoring the items on a four-point scale provides significant benefits in terms of its range of reliable measurement as compared to using a two-point scale.

References

Austin, E. J. (2005). Personality correlates of the broader autism phenotype as assessed by the Autism Spectrum Quotient (AQ). *Personality and Individual Differences*, *38*, 451-460.

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*, 241-251.

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

Booth, T., Murray, A. L., McKenzie, K., Kuenssberg, R., O'Donnell, M., & Burnett, H. (2013). Brief report: An evaluation of the AQ-10 as a brief screening instrument for ASD in adults. *Journal of Autism and Developmental Disorders*,*43*, 2997-3000.

Caspi, A., & Moffitt, T. E. (2006). Gene–environment interactions in psychiatry: joining forces with neuroscience. *Nature Reviews Neuroscience*, *7*, 583-590.

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*,*22*, 265-289.

Constantino J. N. & Gruber C. P. (2005). *Social Responsiveness Scale: Manual*. Los Angeles, CA: Western Psychological Services.

Embretson, S. E. (1996). Item response theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement*, *20*, 201-212.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Psychology Press.

Frazier, T. W., Youngstrom, E. A., Sinclair, L., Kubu, C. S., Law, P., Rezai, A., ... & Eng, C. (2010). Autism spectrum disorders as a qualitatively distinct category from typical behavior in a large, clinically ascertained sample. *Assessment*, *17*, 308-320.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., ... & Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, *31*, 4-19.

Hernández, A., Drasgow, F., & González-Romá, V. (2004). Investigating the functioning of a middle category by means of a mixed-measurement model.*Journal of Applied Psychology*, *89*, 687-699.

Hoekstra, R. A., Bartels, M., Cath, D. C., Boomsma, D. I. (2008). Factor structure, reliability and criterion validity of the Autism-Spectrum Quotient (AQ): A study in Dutch population and patient groups. *Journal of Autism and Developmental Disorders, 38,* 1555-1566.

Hoekstra, R. A., Vinkhuyzen, A. A., Wheelwright, S., Bartels, M., Boomsma, D. I., Baron-Cohen, S., ... & van der Sluis, S. (2011). The construction and validation of an abridged version of the autism-spectrum quotient (AQ-Short).*Journal of Autism and Developmental Disorders*, *41*, 589-596.

Hurley, R. S., Losh, M., Parlier, M., Reznick, J. S., & Piven, J. (2007). The broad autism phenotype questionnaire. *Journal of Autism and Developmental Disorders*, *37*, 1679-1690.

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*, 1706-1714.

Kang, S. M., & Waller, N. G. (2005). Moderated multiple regression, spurious interaction effects, and IRT. *Applied Psychological Measurement*, *29*, 87-105.

Klin, A., Danovitch, J. H., Merz, A. B., & Volkmar, F. R. (2007). Circumscribed interests in higher functioning individuals with autism spectrum disorders: An exploratory study. *Research and Practice for Persons with Severe Disabilities*,*32*, 89-100.

Kuenssberg, R., Murray, A. L., Booth, T., & McKenzie, K. (2014). Structural validation of the abridged Autism Spectrum Quotient–Short Form in a clinical sample of people with autism spectrum disorders. *Autism*, *18*, 69-75.

Kulas, J. T., & Stachowski, A. A. (2009). Middle category endorsement in odd-numbered Likert response scales: Associated item characteristics, cognitive demands, and preferred meanings. *Journal of Research in Personality*, *43*, 489-493.

Kulas, J. T., & Stachowski, A. A. (2013). Respondent rationale for neither agreeing nor disagreeing: Person and item contributors to middle category endorsement intent on Likert personality indicators. *Journal of Research in Personality*, *47*, 254-262.

Kulas, J. T., Stachowski, A. A., & Haynes, B. A. (2008). Middle response functioning in Likert-responses to personality items. *Journal of Business and Psychology*, *22*, 251-259.

Lai, M. C., Lombardo, M. V., Auyeung, B., Chakrabarti, B., & Baron-Cohen, S. (2015). Sex/Gender Differences and Autism: Setting the Scene for Future Research. *Journal of the American Academy of Child & Adolescent Psychiatry*,*54*(1), 11-24.

Lau, W. Y. P., Kelly, A. B., & Peterson, C. C. (2013). Further evidence on the factorial structure of the Autism Spectrum Quotient (AQ) for adults with and without a clinical diagnosis of autism. *Journal of Autism and Developmental Disorders*, *43*, 2807-2815.

Liss, M., Mailloux, J., & Erchull, M. J. (2008). The relationships between sensory processing sensitivity, alexithymia, autism, depression, and anxiety. *Personality and Individual Differences*, *45*, 255-259.

Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *4*, 73-79.

Lundström, S., Chang, Z., Råstam, M., Gillberg, C., Larsson, H., Anckarsäter, H., & Lichtenstein, P. (2012). Autism spectrum disorders and autistic-like traits: similar etiology in the extreme end and the normal variation. *Archives of General Psychiatry*, *69*, 46-52.

Meijer, R. R., & Egberink, I. J. (2012). Investigating invariant item ordering in personality and clinical scales: Some empirical findings and a discussion.*Educational and Psychological Measurement*, *72*, 589-607.

Murray, A. L., Booth, T., McKenzie, K., Kuenssberg, R., & O'Donnell, M. (2014). Are Autistic traits measured equivalently in individuals with and without an Autism Spectrum Disorder? An invariance analysis of the Autism Spectrum Quotient Short Form. *Journal of Autism and Developmental Disorders*, *44*, 55-64.

Murray, A. L., Allison, C., Smith, P., Baron-Cohen, S., Booth, T., Auyeung, B. (2015). Investigating diagnostic bias in autism spectrum conditions: An item response theory analysis of sex bias in the AQ-10. Submitted.

Murray, A. L., Booth, T., & Molenaar, D. (2015). When 'middle' really means 'top' or 'bottom'. An analysis of the 16PF5 using Bock's nomimal response model. Submitted.

Murray, A. L., Kuenssberg, R., McKenzie, K., Booth, T. (2015). Can the Autism Spectrum Quotient (AQ) and its short form (AQ-S) be used to measure general ASD traits, specific ASD traits or both? Submitted.

Murray, A. L., McKenzie, K., Kuenssberg, R., & O'Donnell, M. (2014). Are We Under-Estimating the Association Between Autism Symptoms?: The Importance of Considering Simultaneous Selection When Using Samples of Individuals Who Meet Diagnostic Criteria for an Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, 1-10.

Murray, A. L., Molenaar, D., Johnson, W., & Krueger, R. (2015). Dependence of gene-by-environment interactions (GxE) on scaling: Comparing the use of sum scores, transformed sum scores and IRT scores for the phenotype in tests of GxE. Submitted.

Muthén, L. K., & Muthén, B. O. (1998-2013). Mplus User's Guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

Naito, K, Matsui, Y., Maeda, K., & Tanaka, K. (2010). Evaluation of the validity of the Autism Spectrum Quotient (AQ) in differentiating high-functioning autistic spectrum disorder from schizophrenia. *Kobe Journal of Medical Science*, *56*, 116-24.

Rammstedt, B., & Krebs, D. (2007). Does response scale format affect the answering of personality scales? Assessing the Big Five dimensions of personality with different response scales in a dependent sample. *European Journal of Psychological Assessment*, *23*, 32-38.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement*,*15*, 361-373.

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*, 27-48.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, *16*, 19-31.

Ruzich, E., Allison, C., Smith, P., Watson, P., Auyeung, B., Ring, H. & Baron-Cohen, S. (2015). Measuring autistic traits in the general population: a systematic review of the Autism-Spectrum Quotient (AQ) in a nonclinical population sample of 6,900 typical adult makes and females. *Molecular Autism, 6,* 1-12.

Schwabe, I., & van den Berg, S. M. (2014). Assessing genotype by environment interaction in case of heterogeneous measurement error. *Behavior Genetics*, 1-13.

Sizoo, B. B., van den Brink, W., Gorissen-van Eenige, M., Koeter, M. W., van Wijngaarden-Cremers, P. J., & van der Gaag, R. J. (2009). Using the Autism-Spectrum Quotient to discriminate Autism Spectrum Disorder from ADHD in adult patients with and without comorbid substance use disorder. *Journal of Autism and Developmental Disorders*, *39*, 1291-1297.

Spek, A. A., & Wouters, S. G. (2010). Autism and schizophrenia in high functioning adults: Behavioral differences and overlap. *Research in Autism Spectrum Disorders*, *4*, 709-717.

Thomas, M. L. (2011). The value of item response theory in clinical assessment: A review. *Assessment*, *18*, 291-307.

Van den Oord, E. J., Pickles, A., & Waldman, I. D. (2003). Normal variation and abnormality: an empirical study of the liability distributions underlying depression and delinquency. *Journal of Child Psychology and Psychiatry*, *44*, 180-192.

Wakabayashi, A., Baron-Cohen, S., & Wheelwright, S. (2006). Are autistic traits an independent personality dimension? A study of the Autism-Spectrum Quotient (AQ) and the NEO-PI-R. *Personality and Individual Differences*, *41*, 873-883.

Wheelwright, S., Auyeung, B., Allison, C., & Baron-Cohen, S. (2010). Research defining the broader, medium and narrow autism phenotype among parents using the Autism Spectrum Quotient (AQ). *Molecular Autism*, *1*, 1-10.

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.

Published as: Murray, A. L., Booth, T., McKenzie, K., & Kuenssberg, R. (2015). What Range of Trait Levels Can the Autism-Spectrum Quotient (AQ) Measure Reliably? An Item Response Theory Analysis. *Psychological Assessment,* Online First.

Tables

Table 1:

IRT parameters from bi-factor model with normal ogive link function with AQ items scored on 2-point scale

Item	Abbreviated content	General Factor Slope	Specific Factor Slope	Threshold
		Social Skills Subscale		
1	Prefer to do things on own	0.65	0.52	-0.06
11	Social situations easy	2.18	0.66	-0.49
13	Prefer library to party	0.89	0.50	-0.14
15	Drawn to people versus things	1.12	0.32	0.04
22	Hard to make new friends	1.36	0.08	-0.30
36	Infer thoughts/feelings from faces	1.27	-0.30	0.59
44	Enjoy social occasions	3.29	2.44	1.02
45	Difficult to infer intentions	1.33	-0.51	0.18
47	Enjoy meeting new people	1.48	0.80	0.54

48	Good diplomat	0.94	-0.05	0.39
	Attention Switching Subscale			
2	Prefer to do things the same way	0.71	0.66	-0.47
4	Get so absorbed	0.73	-0.01	-0.84
10	Keep track of several conversations	1.02	-0.30	-0.08
16	Very strong interests	0.56	0.30	-0.41
25	Not upset by daily routine disruption	0.74	0.67	-0.08
32	Easy to multi-task	0.79	-0.16	0.37
34	Enjoy doing things spontaneously	1.07	0.53	0.42
37	Can switch back after interruption	0.73	0.06	0.27
43	Like to plan activities carefully	0.73	0.76	-0.67
46	Anxiety in new situations	0.95	0.42	-1.04
	Attention to Detail Subscale			
5	Notice small sounds	0.29	0.24	-0.53
6	Notice information e.g. number plates	0.39	0.83	-0.16

9	Fascinated by dates	0.34	0.93	0.75
12	Notice details others don't	0.21	0.72	-1.11
19	Fascinated by numbers	0.43	0.91	0.42
23	Notice patterns in things all the time	0.62	1.15	-0.59
28	Concentrate on whole picture vs details	0.64	0.23	-0.01
29	Poor at remembering phone numbers	0.05	0.55	0.01
30	Don't notice small changes	-0.21	0.28	-0.49
49	Not good at remembering dates of birth	-0.03	0.48	0.05
	Communication Subscale			
7	Inadvertently impolite	0.84	0.74	0.60
17	Enjoy social chit-chat	1.09	-0.32	0.09
18	Others can't get word in edgeways	0.18	0.53	0.32
26	Can't maintain conversation	1.42	0.00	-0.33
27	Easy to read between the lines	1.11	0.12	0.49
31	Can't tell if someone is bored listening	0.88	0.16	0.80

33	Turn-taking talking on the phone	0.75	0.26	0.35
35	Last to understand point of joke	0.75	0.56	0.46
38	Good at social chit-chat	1.74	-	-0.26
39	Go on about the same thing	1.20	1.33	0.21
	Imagination Subscale			
3	Hard to create picture in mind	0.58	0.86	1.17
8	Easily imagine characters in a story	0.80	0.63	0.88
14	Making up stories is easy	0.37	0.56	0.28
20	Work out character intentions in story	0.95	0.46	0.85
21	Don't enjoy fiction	0.52	0.45	0.77
24	Prefer theatre to museum	0.35	0.05	0.04
40	Enjoyed pretend play as a child	0.77	0.32	0.61
41	Collect information about categories	0.67	-0.08	0.32
42	Difficult to imagine being someone else	0.86	0.23	0.23
50	Can play pretend with children	0.84	0.54	0.32

Published as: Murray, A. L., Booth, T., McKenzie, K., & Kuenssberg, R. (2015). What Range of Trait Levels Can the Autism-Spectrum Quotient (AQ) Measure Reliably? An Item Response Theory Analysis. *Psychological Assessment,* Online First.

Table 2:

IRT parameters from bi-factor model with normal ogive link function with AQ items scored on 4-point scale

Item	General Factor Slope	Specific Factor Slope	Threshold 1	Threshold 2	Threshold 3
		Social Skills Subscale			
1	0.57	0.53	-1.29	-0.10	1.05
11	1.94	0.65	-2.01	-0.52	0.89
13	0.83	0.68	-1.21	-0.15	0.78
15	1.11	0.35	-1.21	0.00	1.40
22	1.42	0.19	-1.40	-0.34	0.67
36	1.15	-0.29	-0.96	0.51	1.40
44	2.26	1.74	-1.82	0.66	2.61
45	1.34	-0.39	-1.49	0.12	1.51
47	1.52	0.95	-1.51	0.52	1.98
48	0.85	-0.08	-1.03	0.33	1.29
		Attention Switching Subscale			

2	0.83	0.74	-1.76	-0.52	0.93
4	0.70	0.13	-1.80	-0.85	0.30
10	1.04	-0.11	-1.39	-0.11	0.82
16	0.53	0.37	-1.60	-0.41	0.62
25	0.69	0.66	-1.24	-0.11	1.02
32	0.76	-0.03	-0.93	0.33	1.10
34	0.96	0.41	-1.02	0.34	1.32
37	0.75	0.09	-1.11	0.24	1.27
43	0.67	0.59	-1.73	-0.62	0.67
46	0.96	0.33	-2.04	-1.02	0.27
	Attention to Detail Subscale				
5	0.29	0.29	-1.30	-0.55	0.45
6	0.41	0.95	-1.04	-0.18	0.96
9	0.34	0.78	-0.23	0.69	1.51
12	0.27	0.68	-2.29	-1.10	0.34

19	0.38	0.85	-0.50	0.38	1.39
23	0.48	1.03	-1.72	-0.56	0.85
28	0.70	0.29	-1.31	-0.05	1.24
29	0.08	0.48	-0.78	0.00	1.03
30	-0.11	0.25	-1.29	-0.48	0.74
49	-0.06	0.39	-0.71	0.05	0.91
Communication Subscale					
7	0.86	0.51	-0.29	0.53	1.48
17	1.23	-0.83	-1.44	0.09	1.14
18	0.21	0.44	-0.58	0.30	1.26
26	1.40	-0.23	-1.58	-0.36	0.98
27	1.08	0.23	-0.95	0.44	1.36
31	0.80	0.19	-0.54	0.75	1.41
33	0.72	0.13	-0.82	0.32	1.25
35	0.77	0.36	-0.64	0.43	1.35

38	3.17	-1.82	-3.27	-0.47	1.90
39	0.95	0.72	-1.01	0.12	1.30
		Imagination Subscale			
3	0.47	0.56	-0.16	1.00	1.75
8	0.74	0.65	-0.32	0.84	1.84
14	0.34	0.64	-0.82	0.27	1.06
20	0.77	0.31	-0.44	0.76	1.52
21	0.42	0.46	0.06	0.75	1.25
24	0.44	0.05	-0.87	0.02	0.94
40	0.69	0.44	-0.45	0.59	1.27
41	0.68	-0.07	-0.44	0.29	1.09
42	0.81	0.26	-0.98	0.20	0.99
50	0.87	0.55	-0.96	0.30	1.06

Note. See Table 1 for abbreviated item contents.

Published as: Murray, A. L., Booth, T., McKenzie, K., & Kuenssberg, R. (2015). What Range of Trait Levels Can the Autism-Spectrum Quotient (AQ) Measure Reliably? An Item Response Theory Analysis. *Psychological Assessment,* Online First.

Figure Captions

Figure 1

Conditional test information function for the general factor with items scored on a 2-point scale

Figure 2

Conditional test information function for the general factor with items scored on a 4-point scale