RESEARCH ARTICLE

A Comprehensive Psychometric Analysis of Autism-Spectrum Quotient Factor Models Using Two Large Samples: Model Recommendations and the Influence of Divergent Traits on Total-Scale Scores

Michael C. W. English, Gilles E. Gignac, Troy A. W. Visser, Andrew J. O. Whitehouse, and Murray T. Maybery

The Autism-Spectrum Quotient (AQ) is a psychometric scale that is commonly used to assess autistic-like traits and behaviors expressed by neurotypical individuals. A potential strength of the AQ is that it provides subscale scores that are specific to certain dimensions associated with autism such as social difficulty and restricted interests. However, multiple psychometric evaluations of the AQ have led to substantial disagreement as to how many factors exist in the scale, and how these factors are defined. These challenges have been exacerbated by limitations in study designs, such as insufficient sample sizes as well as a reliance on Pearson, rather than polychoric, correlations. In addition, several proposed models of the AQ suggest that some factors are uncorrelated, or negatively correlated, which has ramifications for whether total-scale scores are meaningfully interpretable—an issue not raised by previous work. The aims of the current study were to provide: (a) guidance as to which models of the AQ are viable for research purposes, and (b) evidence as to whether total-scale scores are adequately interpretable for research purposes. We conducted a comprehensive series of confirmatory factor analyses on 11 competing AQ models using two large samples drawn from an undergraduate population (*n* = 1,702) and the general population (*n* = 1,280). Psychometric evidence largely supported using the three-factor model described by Russell-Smith et al. [Personality and Individual Differences 51(2), 128–132 (2011)], but did not support the use of total-scale scores. We recommend that researchers consider using AQ subscale scores instead of total-scale scores. ***Autism Res*** *2020, 13: 45–60.* © 2019 International Society for Autism Research, Wiley Periodicals, Inc.

Lay Summary: We examined 11 different ways of scoring subscales in the popular Autism-Spectrum Quotient (AQ) questionnaire in two large samples of participants (i.e., general population and undergraduate students). We found that a three-subscale model that used "Social Skill," "Patterns/Details," and "Communication/Mindreading" subscales was the best way to examine specific types of autistic traits in the AQ. We also found some weak associations between the three subscales—for example, being high on the "Patterns/Details" subscale was not predictive of scores on the other subscales. This means that meaningful interpretation of overall scores on the AQ is limited.

Keywords: autism-spectrum quotient; AQ; confirmatory factor analysis; Raine study; autism; autistic traits; factor model

Introduction

It is well established that many of the traits and behaviors associated with Autism Spectrum Disorder (ASD) are not restricted to those with clinical diagnoses. While this observation was initially noted in the relatives of autistic individuals [e.g., Bishop, Maybery, Wong, Maley, & Hallmayer, 2006], subsequent work has robustly demonstrated similar autistic-like traits in the broader neurotypical population [Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001; Constantino & Todd, 2003]. Indeed, the distribution of autistic-like traits in the general population is smooth in extending out to the clinical extreme [e.g., Skuse, Mandy, & Scourfield, 2005] and twin studies suggest similar etiology for ASD and autistic traits in the general population [for review, see Ronald & Hoekstra, 2011]. Critically, individuals with high levels of autistic-like traits are often used as a proxy for individuals with clinically diagnosed autism, with studies generally conducted in one of two ways. One is by comparing individuals selected for relatively low and high levels of autistic-like traits on other measures of interest, which mimics studies comparing non-autistic and autistic individuals, while the other is by treating trait level as a continuous dimension that can be related to other continuous measures. The benefits of using neurotypical participants in such a manner for autism research include the relative ease

From the School of Psychological Science, University of Western Australia, Perth, Western Australia, Australia (M.C.W.E., G.E.G., T.A.W.V., M.T.M.); Telethon Kids Institute, Perth, Western Australia, Australia (A.J.O.W.)

Received March 25, 2019; accepted for publication August 9, 2019

Address for correspondence and reprints: Michael C. W. English, School of Psychological Science, University of Western Australia, Perth, Western Australia, Australia. E-mail: michael.english@uwa.edu.au

Published online 29 August 2019 in Wiley Online Library (wileyonlinelibrary.com)
DOI: 10.1002/aur.2198

of participant recruitment, enabling larger samples, and the ability to easily control for potential confounds, such as IQ differences [Landry & Chouinard, 2016].

Just as individuals diagnosed with autism are increasingly recognized as a highly heterogeneous group (with respect to, e.g., genetics) [Geschwind & State, 2015], behavior [Fountain, Winter, & Bearman, 2012; Lord, Bishop, & Anderson, 2015; Pickles, Anderson, & Lord, 2014], and cognition [Feczko et al., 2018; Taylor, Maybery, Grayndler, & Whitehouse, 2014], so too is heterogeneity expressed in the display of autistic-like traits in neurotypical individuals. Popular measures of autistic-like traits, such as the Social Responsiveness Scale-2 [SRS-2; Constantino & Gruber, 2012] and Autism-Spectrum Quotient [AQ; Baron-Cohen et al., 2001], can be scored on specific trait dimensions [e.g., *Social Avoidance* or *Insistence on Sameness* for the SRS-2; Frazier et al., 2014] to provide additional nuance to total scale scores. Dimensions of autistic traits have been shown to be separable in heritability [e.g., Ronald et al., 2006], cognitive features [e.g., Russell-Smith, Maybery, Bayliss & Sng, 2012] and patterns of behavior [e.g., Davis et al., 2017]. While the ability to test for such relationships with this level of detail is necessary to fully understand the heterogeneous nature of autism, researcher's investigations will only be as good as the quality and understanding of the psychometric scales that they use [Frazier et al., 2014].

Autism-Spectrum Quotient

The AQ is an interesting case. Despite being one of the mostly widely used self-report questionnaires of autistic-like traits, with over 2,300 citations on Scopus (as of June 10, 2019), there is little consensus as to how many individual trait dimensions can be measured with the scale or which specific items define these dimensions. In addition to Total-Scale AQ scores, the 50-item scale was originally described to consist of five subscales, each theoretically derived to tap into specific aspects associated with autism: *Social Skill*, *Attention Switching*, *Attention to Detail*, *Communication*, and *Imagination* [Baron-Cohen et al., 2001]. The authors demonstrated that subscales had adequate internal consistency (range 0.63–0.77), suggesting sound psychometric properties and justifying the use of the AQ for research purposes. Moreover, given that clinical ASD is increasingly recognized as being multidimensional in nature, with clinical presentations demonstrating substantial variability in the endorsement of the different diagnostic criteria [see, e.g. Kim, Macari, Koller, & Chawarska, 2016; Mandy & Skuse, 2008; Whitten, Unruh, Shafer, & Bodfish, 2018], the ability of the AQ to also tap into separate trait dimensions can be considered a considerable strength of the scale.

Factor Analyses of the AQ

Following the development of the AQ, other researchers have independently examined the psychometric properties of the scale and its original factor structure. On one hand, a number of studies show that total scale AQ scores appear to be associated with respectable internal consistency, as measured using coefficient (Cronbach's) alpha (see Table 1). On the other hand, issues have been reported with respect to the factor structure of the AQ. First, while multiple attempts have been made to verify the original subscale structure of the AQ, none of the three published confirmatory factor analysis (CFA) studies we identified confirmed the five-factor model [Hoekstra et al., 2008; Kloosterman, Keefer, Kelley, Summerfeldt, & Parker, 2011; Lau, Gau, et al., 2013]. Second, other work has found generally inadequate internal consistency associated with the original factors (see Table 1), suggesting that these factors are poorly defined and are of questionable interpretability.

A surface-level interpretation of these findings would likely lead one to the conclusion that, while caution should be taken when attempting to make inferences using the individual subscales of the original model, the overall scale is psychometrically sound and total scale AQ scores are interpretable for research purposes. However, as we discuss in further detail below, respectable internal consistency for the total scale AQ score is not necessarily

Table 1. Coefficient Alphas Calculated for Each Component in the Original Five-Component Model of the Autism-Spectrum Quotient across Separate Data Sets

Study	Total scale	Social skills	Attention switching	Attention to details	Communication	Imagination
Baron-Cohen et al. [2001]	Not given	0.77	0.67	0.63	0.65	0.65
Austin [2005]	0.82	0.75	0.58	0.66	0.61	0.65
Hurst et al. [2007]	0.67	0.66	0.41	0.60	0.47	0.40
Hoekstra, Bartels, Cath, and Boomsma [2008][a]						
Undergraduates	0.81	0.76	0.63	0.63	0.52	0.63
General population	0.71	0.69	0.62	0.68	0.49	0.52
Freeth, Sheppard, Ramachandran, and Milne [2013] [UK]	0.79	0.63	0.56	0.56	0.59	0.53
Lau et al. [2013][a]	Not given	0.56	0.42	0.45	0.49	0.39

[a] Hoekstra et al. [2008] and Lau, Gau et al. [2013] used Dutch and Chinese translations of the Autism-Spectrum Quotient, respectively.

meaningful, in combination with relatively small positive correlations between the AQ factors/subscales. Unfortunately, such correlations have been rarely reported in the previous work. Furthermore, internal consistency cannot be relied upon as an index of homogeneity or dimensionality [Hattie, 1985]. Consequently, testing for the presence of truly homogenous (i.e., correlated) traits is needed to determine whether it is appropriate to be using and interpreting total scale AQ scores, and factor analysis is required to empirically identify the dimensions associated with a multi-item scale [Reise, Waller, & Comrey, 2000].

Alternative Models of the AQ: Inconsistency and Disagreement

Beginning with Austin [2005], substantial efforts have been made to identify more psychometrically sound factor structures in the AQ. However, as can be seen in Figure 1, we identified 11 different factor models of the AQ endorsed in the literature, and those factor models show considerable heterogeneity. Specifically, the number of factors identified range from two to five. Furthermore, two studies proposed relatively more complex models [i.e., four factors grouped together under a hierarchical factor; Hoekstra et al., 2008, 2011]. The factors also vary in nature. For instance, while socially oriented and details/patterns-oriented factors were always present, an imagination factor was inconsistently identified.

There is also substantial disagreement observed at the item level (see Fig. 1). Specifically, when clustering together notionally similar factor labels across the different models, only 7 of the 50 items always load on a factor of a comparable type across all the models observed (items 6, 11, 19, 22, 23, 44, and 47). Fourteen items are observed to inconsistently load on a factor of a comparable type and to not load on any factor in some instances (items 1, 3, 5, 9, 12, 13, 14, 15, 16, 27, 29, 31, 33, and 35). Twenty-nine items load on at least two dissimilar factors (items 2, 4, 7, 8, 10, 16, 17, 18, 20, 21, 25, 26, 28, 30, 32, 34, 36, 37, 38, 39, 40, 41, 42, 43, 45, 46, 48, 49, and 50). Eight items never load on a factor comparable to the component to which they were assigned in the original model (items 17, 26, 30, 36, 38, 41, 46, and 49). Finally, one item (item 24: "I would rather go to a party than a museum") never loads on any factor in any of the models.

Inconsistencies have also been reported in the correlations between the estimated factors across the different models. This is an important issue, because evidence for substantially divergent factors (i.e., essentially uncorrelated) would suggest that total scale AQ scores are not inherently interpretable. That is, different people could show identical AQ scores but have substantially different underlying factor scores. Based on our review, while some factor analytic studies have reported a pattern of mostly positive correlations between the identified AQ factors [e.g., Lau, Kelly, & Peterson, 2013 found significant, positive correlations between all five of their factors], other studies have reported at least one negative and/or non-significant correlation between AQ dimensions [Austin, 2005; Kloosterman et al., 2011; Lau, Gau, et al., 2013; Russell-Smith, Maybery, & Bayliss, 2011; Stewart & Austin, 2009], undermining the interpretability of total scale AQ scores.

Limitations in the Development of Alternative Factor Structures

Though the factor analytic studies reviewed above benefited from using a rigorous quantitative approach to derive a factorial model of the AQ, these studies also had substantial limitations. For example, the vast majority of the studies did not conduct polychoric correlations [Holgado-Tello, Chacón-Moscoso, Barbero-García, & Vila-Abad, 2010], which are more appropriate for items measured with Likert scale [the exceptions were Hoekstra et al., 2011, Hoekstra et al., 2008]. A polychoric correlation estimates an association between ordinal data more accurately than does a Pearson correlation, which tends to underestimate the association for ordinal variables, particularly when the items are scaled on a Likert scale with less than five points [Asún, Rdz-Navarro, & Alvarado, 2016; but see DiStefano, 2002, for evidence that even a 5-point Likert scale may be problematic for the Pearson correlation, if the data are substantially non-normally distributed]. Importantly, the AQ items are scaled on a 4-point Likert scale [Baron-Cohen et al., 2001].

Additionally, it appears that previously published studies lacked adequate sample sizes. While there are numerous recommendations for minimum sample sizes for conducting exploratory factor analyses [see Costello & Osborne, 2005], a review of the literature suggests that accurate and stable factor solutions are dependent upon the level of communality associated with the indicators [Hogarty, Hines, Kromrey, Ferron, & Mumford, 2005]. In the context of item-level factor analyses, it would be expected that indicator communality would be relatively low (say, <0.10), thus, larger sample sizes would be required. For example, Hirschfeld, von Brachel, and Thielsch [2014] found that a stable factor structure for a 42-item personality inventory was not achieved until $N \approx 1{,}000$. Consequently, if we were to consider the minimum number of participants required to adequately examine the factor structure of the 50-item AQ as at least $N \approx 1{,}000$, for many previous studies, samples sizes are arguably inadequate [though exceptions include Hoekstra et al., 2011; Lau, Gau, et al., 2013]. This alone might also explain the lack of consistent factor structures reported across these studies [Hirschfeld et al., 2014].

Finally, it is possible that the generalizability of the factors defined in many studies may be limited because they were based on data collected solely from undergraduate students, rather than the general population [Austin, 2005; Freeth et al., 2013; Kloosterman et al., 2011; Russell-Smith

Model	AQ1	AQ2*	AQ3	AQ4*	AQ5*	AQ6*	AQ7*	AQ8	AQ9*	AQ10	AQ11	AQ12*	AQ13*	AQ14	AQ15	AQ16*	AQ17	AQ18*	AQ19*	AQ20*	AQ21*	AQ22*	AQ23*	AQ24	AQ25
1	A	E^1	D	E^1	C^1	C^1	B	D	C^1	E^1	A	C^1	A	D	A	E^1	B	B	C^1	D	D	A	C^1	D	E^1
2					C^1	C^1	B		C^1		A	C^1	A		A		A		C^1	B		A	C^1		C^1
3	A	B	D	D	C^1	C^1	B	D	C^1	B	A	C^1	A	D	A		A	A	C^1	B	B	A	C^1		
4	A				C^1	C^1			C^1	A	A	C^1	A		A		A		C^1	B		A	C^1		
5	A		D	D	C^1	C^1		D	C^1	B	A	C^1	A	D	A	C^1	A		C^1	B	D	A	C^1		
6	A	E^1	D	E^1	C^1	C^1		D		B	A	C^1	A		A		A	E^1	C^1	D	D	A	C^1		E^1
7	A	E^2	D	E^1		C^1		D	C^1	E^1	A		A	D	A				C^1	D		A	C^1		E^2
8	A		D		C^1	C^1	B	D	C^1	B	A	C^1	A	D	A	C^1	A		C^1	B		A	C^1		
9	A			E^1	C^1	C^2	B		C^2	A	A	C^1	A			E^1	A		C^2	B		A	C^1		
10	A	E^1		C^2	C^2	C^1	C^2	B	C^1	B	A	C^2	A		A	C^2	A		C^1	B		A	C^2		E^1

Model	AQ26*	AQ27	AQ28	AQ29	AQ30	AQ31	AQ32	AQ33*	AQ34	AQ35*	AQ36	AQ37	AQ38	AQ39*	AQ40	AQ41*	AQ42*	AQ43*	AQ44	AQ45*	AQ46*	AQ47	AQ48	AQ49	AQ50
1	B	B	C^1	C^1	C^1	B	E^1	B	E^1	B	A	E^1	B	B	D	D	D	E^1	A	A	E^1	A	A	C^1	D
2	A								A	B		B	A	B	A			C^1	A	B		A			A
3	A	B		C^1	B	B	B	B		B	B	B	A	B	D	C^1			A	B	A	A	B	D	D
4	A	B				B			A	B	B		A	B		C^1			A	B	A	A	B		
5	A	B		C^1	B	B	B		A	B	B	B	A		D	C^1			A	B	A	A			D
6		B				B					B		A	E^1	D				A	B		A			
7							E^1		E^2			E^1				C^1	D		A	A	A	A			D
8	A	B			B			B	D		B		A			C^1	B		A	B	A	A	B		D
9	A	B	C^1	C^2		B	E^1	B	E^1	B	B	E^1	A	E^1		C^2			A	B	A	A	A		
10	A	B	B			B	B		E^1		B	B	A	C^2		C^1	B	E^1	A	B	A	A	A		A

Models		Types of Factors	
1. Baron-Cohen et al (2001) & Hoekstra et al. (2008)	6. Kloosterman et al. (2011)	A	Social Factors
2. Austin (2005)	7. Hoekstra et al. (2011)	B	Communications Factors
3. Stewart & Austin (2009)	8. Freeth et al. (2013) [UK]	C	Attention, Patterns & Details Factors
4. Russell-Smith et al. (2011) [3-factors]	9. Lau, Gau, et al. (2013)	D	Imagination Factors
5. Russell-Smith et al. (2011) [4-factors]	10. Lau, Kelly & Peterson (2013)	E	Rep. Behaviour & Routines Factors
			Item not assigned a factor

Figure 1. A summary of the factors and item loadings in each of the 11 models examined, with conceptually similar factors grouped together. Note that Model 1 represents both the original Baron-Cohen [2001] model and an identical variant with an additional hierarchical factor [Hoekstra et al., 2008]. Asterisk denotes reverse-keyed items. Superscript numbers distinguish conceptually similar, but separate, factors in a single model. Not shown are hierarchical factors described in two models [Hoekstra et al., 2008, 2011].

et al., 2011; Stewart & Austin, 2009]. While the factor solutions reported in these studies may be adequate for an undergraduate student population, it is not necessarily the case that such factor solutions would generalize to the general population [Hanel & Vione, 2016; Henrich, Heine, & Norenzayan, 2010; Peterson & Merunka, 2014].

The Present Study

Given the increase in research investigating the heterogeneous nature of autism, it is likely that autism researchers using the AQ will be increasingly likely to use AQ factor scores, in addition to total scale scores, in their research designs. However, such investigations will be hindered if psychometrically unsound and poorly defined factors are used to examine specific autistic-like trait dimensions [e.g., see Ronald and Hoekstra [2011] for a discussion of the influence of measurement error on estimating trait heritability]. As we have outlined earlier, not only has an expanding body of psychometric work questioned the empirical validity of the original factor structure of the AQ, there is little consensus among empirically derived, alternative factor models in terms of the number of factors the item make-up of these factors, and the direction and strength of the associations between the factors. This last point is especially critical, as the nature and strength of the correlations between the factors largely determines whether total scale AQ scores are justifiably interpretable or not [Gignac, 2014]. Consequently, it is important for researchers to have some confidence with respect to which factor model of the AQ is the most appropriate for research purposes, as such knowledge informs the method to score the questionnaire, not to mention help us understand the nature of the autism spectrum.

Consequently, the first aim of the present study is to determine which factor model of the AQ (if any) researchers should be choosing to implement in their study designs. This would be achieved by conducting a comprehensive series of confirmatory factor analyses across a range of competing factor models of the AQ. Importantly, these analyses seek to overcome many of the methodological limitations outlined above. Critically, our analyses will involve examining two particularly large samples drawn from undergraduate students and the general population (each $N > 1{,}000$). Not only does this address previous issue with inadequate sample sizes, but also allows us to statistically determine if it is appropriate to apply a single AQ factor model to both populations. Furthermore, we used the more appropriate polychoric correlations, rather than Pearson correlations.

While it would be arguably simpler to conduct a new exploratory factor analysis with this methodology and our data sets, and promote the resulting model as "better" than prior models, we chose to conduct a comprehensive series of confirmatory analyses for two reasons. First, such an extensive series of analyses will, hopefully, make a strong case for those models that do and do not have psychometric support. Second, we feel that this body of research is crowded enough without another model of the AQ being proposed. Given that one of the existing models might be shown to have strong psychometric support in our analyses, we elected to only develop another alternative model should none of the existing AQ models gain sufficient psychometric support.

The second aim of the present study was to determine the degree to which total scale AQ scores are interpretable by assessing the inter-dimension correlations (i.e., the homogeneity of AQ factors) of the "best" model from the previous analyses. Calculating internal consistency in this manner is preferable for multidimensional scales, as it is less likely to overestimate internal consistency, an effect shown to occur when calculating internal consistency at the commonly used item unit-level [Gignac, 2014].

Method

Participants

Participant data were obtained across several studies, all of which were approved by the relevant human research and ethics committees at the University of Western Australia, King Edward Memorial Hospital, and/or Princess Margaret Hospital for Children in Perth. Informed consent was obtained from all participants prior to data collection.

Undergraduate Sample

Undergraduate data were collected from multiple cohorts of second-year undergraduate students enrolled in Psychology at the University of Western Australia over a consecutive 5-year period spanning 2014–18. Each year, AQ scores were collected from students during a laboratory exercise that was part of their course requirements. The data from 1,702 students who consented to their data being used for research purposes were used for the analyses in the present study. The sample had a mean age of 21.38 years (SD = 6.12; 15 participants missing age data) and included 484 males and 1,218 females. No participant had data excluded from subsequent analyses for any reason.

General Population Sample

General population data were sourced from participants in the Western Australian Pregnancy Cohort (Raine) Study (www.rainestudy.org.au). The Raine Study is a longitudinal investigation based in Western Australia that commenced with recruitment of 2,900 pregnant women between May 1989 and November 1991. Recruitment criteria included a gestational age between 16 and 20 weeks during the recruitment period, English proficiency, and an intention to remain in Western Australia for future follow-ups. At the end of the recruitment period, a sample of 2,868 live births was recorded, and these individuals have been followed-up at regular intervals since. The representativeness of the Raine Study participants has been confirmed at multiple time-points by making comparisons between the participants and nonparticipants of a similar age from Western Australia [Straker et al., 2017]. The present study uses data sourced from the 1,280 Generation 2 individuals (611 male, 669 female) who participated in the age 19–20 years follow-up. No participant had data excluded from subsequent analyses for any reason.

Materials

Autism-spectrum quotient. The AQ [Baron-Cohen et al., 2001] is a 50-item self-report questionnaire designed to assess levels of autistic-like traits and behaviors in neurotypical individuals. The scale includes statements that are intended to tap into autistic-like tendencies. Participants respond using a four-point Likert scale with the responses "Definitely Agree," "Slightly Agree," "Slightly Disagree," and "Definitely Disagree." The original item-level scoring format of the AQ coded responses dichotomously (i.e., into "agreement" or" disagreement"). Responses that endorse the autism phenotype would score one point, regardless of the strength of endorsement, while responses that do not endorse the phenotype would score zero, also regardless of endorsement strength. Consequently, total scale AQ scores could range from 0 to 50, where higher scores indicate more autistic traits. Beginning with Austin [2005], others have adopted a 1–4 scoring strategy, in order to take advantage of the full range of potentially useful information in each item thus increasing scale discriminability [Stevenson & Hart, 2017]. Such a scoring strategy results in total scale AQ scores ranging 50–200. For the current study, we adopted the 1–4 scoring format across all analyses.

Procedure

Statistical analyses. We tested 11 competing correlated-factor models (see Fig. 1) on our undergraduate and general population samples with CFA, using the Lavaan package (0.6–2) developed for R (3.5.1). The 11 models corresponded to the solutions reported in the previously published factor analytic literature on the AQ, as reviewed above. Polychoric correlations were used to estimate the associations between the AQ items. Correspondingly, the models were estimated with weighted least squares estimation. As each of the 11 models was fitted to both the undergraduate and general population cohorts separately, the results are also reported separately.

Consistent with convention [Schweizer, 2010], the CFA models were evaluated using several close-fit indices. Specifically, we consulted two absolute close-fit indices (RMSEA and SRMR; <0.08 is indicative of fair fit; <0.06 is indicative of good fit) and two incremental close-fit indices (CFI and TLI; >0.90 is indicative of fair fit; >0.95 is indicative of good fit). We report the excessively powerful chi-square test statistic for completeness.

In addition to these fit statistics, we also examined two other metrics that, while more commonly used in the initial development of a factor model, are also useful in assessing the quality of a given model. First, we report the percentage of items for each model that had factor loadings above 0.4 following the CFA. This number has been recommended as a cut off for practical interpretation during exploratory factor analysis [Stevens, 2009] and was also used as the minimum factor loading accepted in the development of several previous AQ models [Austin, 2005; Freeth et al., 2013; Kloosterman et al., 2011]. Second, we report coefficient alpha, a measure of internal consistency estimated for the composite (subscale) scores associated with each factor within each model, with higher coefficients indicative of more favorable models. A minimum coefficient of 0.70 is typically required for basic research and practical interpretability [Nunnally & Bernstein, 1994].

In order to assess the utility of total scale AQ scores, we also calculated the inter-factor correlations in each of the 11 models, as well as the internal consistency of the total scale AQ scores, using inter-subtest coefficient alpha [essentially equivalent to omega hierarchical, see Green & Yang, 2015], as recommended by Gignac [2014], since such a method reduces the conflation of specific factor variance with global factor variance. To aid the reader in visualizing the impact of strong or weak factor correlations on total scale AQ scores, we elected to graphically represent the distribution of factor scores displayed by a subset of participants with identical total scale AQ scores. This was achieved using a series of pie-charts, with each chart representing a different participant, and each segment of the chart representing a different factor. To create each pie chart, the factor scores of the participants were adjusted to control for factor size such that the lowest possible score on a factor became "0" and highest possible score became "1" regardless of the number of items on the factor. If identical total scale AQ scores all reflect similar underlying factor scores (high inter-factor correlations) then the pie charts should look very similar. On the other hand, if identical total scale AQ scores reflect different underlying factor scores across individuals, then the pie charts should look dissimilar.

We also conducted a multigroup factorial invariance analysis on the "best fitting" model to determine whether this model was equally applicable to our undergraduate and general population samples. To do this, we assessed measurement invariance between the two samples examined. An individual's score obtained from a test or scale can be considered "measurement invariant" if the likelihood of obtaining that particular score is not dependent on the individual's group membership [Mellenbergh, 1989; Vandenberg & Lance, 2000]. To test this, the difference in the fit of a factor model between two samples is examined using a series of regressions in which additional constraints are imposed on the factor model in incremental steps [for a detailed explanation, see Milfont & Fischer, 2010; Wu, Li, & Zumbo, 2002]. Should the difference in model fit between our undergraduate and general population samples exceed a certain threshold [Chen, 2007; Cheung & Rensvold, 2002], the model would not be considered measurement invariant as an individual's group membership substantially influences their AQ scores. In addition to testing whether the model fit was comparable between the two samples, model fit as a function of sex within each sample was also assessed given previous reports of sex differences in the distribution of autistic-like traits [Abu-Akel, Allison, Baron-Cohen, & Heinke, 2019; Baron-Cohen et al., 2001].

Results

Psychometric Fit of the Competing Models

The results of a series of CFAs in which the identified models were fitted to each of the undergraduate and general population samples separately are described in Table 2. Of particular note, the Russell-Smith et al. [2011] correlated three-factor model (*Social Skills*, *Details/Patterns*, and *Communication/Mindreading*) had the largest number of attractive psychometric properties across both samples, including acceptable model close-fit across all indices (for the student and general population samples respectively, χ^2 = 3,564/2,716; CFI = 0.96/0.94; TLI = 0.95/0.94; RMSEA = 0.07/0.07; SRMR = 0.07/0.08), number of respectable loadings (89% of items in the model loaded >0.40 on their respective factors), and acceptable composite score reliability for the subscale scores (all >0.70, except the *Communication/Mindreading* factor which reached 0.68 and 0.67 for the student and general population samples, respectively). The individual values used to calculate the mean coefficient alphas are reported in Table S1.

By comparison, none of the other models published previously met all of the psychometric criteria, and two of the models did not achieve any of the psychometric criteria [i.e., Hoekstra et al., 2008 and Stewart & Austin, 2009]. In addition, the original AQ factor model [Baron-Cohen et al., 2001] failed to produce adequate measures of statistical fit. In contrast, several alternative models demonstrated reasonable fitness across several indices. It was also noted during the calculation of the average coefficient alpha across each model's factors (see Table S1) that social factors and details/patterns factors ("A" and "C" factors from Table 1 respectively) commonly showed acceptable coefficient alpha values, while the remaining factors never reached acceptable levels.

Inter-Factor Correlations

Inter-factor correlations were calculated separately for each model and sample. As outlined in Table 3, inter-factor correlations in the best-fitting three-factor Russell-Smith et al. [2011] model were consistent across both samples. However, the strength of the correlations varied, with positive correlations observed involving the *Social Skill* dimension, and an absence of correlation observed when the *Social Skill* dimension was not present (see Table 3). For completeness we calculated the inter-factor correlations across all of the models (see Table S2) and found that the strength and direction of the correlations were typically identical between the two samples across all models, thus bolstering the notion that AQ factor structures did not differ across the undergraduate and general population samples.

Comparing Model Fit between Undergraduate and General Population Samples

A multigroup factorial invariance analysis was conducted using the Russell-Smith et al. [2011] correlated three-factor model to statistically examine whether the model fit differed between the undergraduate and general population samples. Generally, configural, metric, and scalar invariance must be achieved before scores from different groups can be meaningfully compared. As can be seen in Table 4, the fit indices did not substantially vary as additional constraints were imposed, indicating that the three-factor model was equally appropriate for both samples.

Finally, additional invariance tests were conducted to determine if sex substantially influenced the model fit. These tests were conducted in an identical manner as the previous measurement invariance analysis except that sex was the grouping variable (instead of sample) and the two samples were examined separately. As detailed in Table 4, each of the examined levels of measurement invariance suggested that the same model comparably fitted data obtained from males and females in each of the samples. The tests of invariance were notably weaker in the general population sample. Though CFI for the configural model dipped below the >0.95 guideline used in our prior analyses [Schweizer, 2010], we chose to provisionally accept the model given that RMSEA was more-than-acceptable. Additionally, subsequent models did not substantially differ, with ΔCFI and ΔRMSEA remaining well within recommended guidelines [Chen, 2007; Cheung & Rensvold, 2002].

Assessing Interpretability of Total Scale AQ Scores

As can be seen in Table 3, the correlations between the factors associated with the Russell-Smith et al. [2011] correlated-factor model varied in magnitude across factor pairs and modest across both samples. Specifically, whereas the *Social Skill* and *Communication/Mindreading* dimensions correlated at $r \approx 0.45$, the *Social Skill* and *Details/Patterns* dimensions correlated at only $r \approx 0.10$, and, importantly,

Table 2. A Summary of Fit Indices and Other Statistics Assessing Factor Structure Quality for Each Factor Structure (Boldface Numbers Are above Fit Thresholds: CFI & TLI > 0.95, RMSEA & SRMR <0.08, 80% items load >0.40, Mean Coefficient Alpha >0.70)

		Baron-Cohen et al. [2001]	Austin [2005]	Hoekstra et al. [2008]	Stewart and Austin [2009]	Russell-Smith et al. [2011] [3-factor]	Russell-Smith et al. [2011] [4-factor]	Kloosterman et al. [2011]	Hoekstra et al. [2011]	Freeth et al. [2013] [UK]	Lau, Gau, et al. [2013]	Lau, Kelly, and Peterson [2013]
	Factors	5	3	2[a]	4	3	4	5	2[a]	4	5	5
	Items	50	26	50	43	28	37	28	28	34	35	39
	df	1,165	296	1,170	854	347	623	340	318	521	550	692
Undergraduate	*Chi-square*[b]	18,102	4,196	18,119	10,745	3,564	6,388	3,260	3,011	6,227	7,152	6,942
sample (n = 1,702)	CFI	0.836	0.938	0.836	0.894	**0.958**	0.931	0.948	0.927	0.933	0.920	0.929
	TLI	0.827	0.932	0.828	0.888	**0.954**	0.926	0.942	0.919	0.928	0.913	0.924
	RMSEA	0.092	0.088	0.092	0.083	**0.074**	**0.074**	**0.071**	**0.071**	**0.080**	0.084	**0.073**
	SRMR	0.095	0.084	0.095	0.085	**0.074**	**0.076**	**0.071**	**0.072**	0.081	0.085	**0.076**
	% Item loadings >0.40	62.00%	76.92%	62.00%	72.09%	**89.29%**	**81.08%**	**82.14%**	**82.14%**	**82.35%**	77.14%	76.92%
	Mean coefficient alpha of factors	0.69	**0.71**	0.68	0.68	**0.77**	0.69	0.67	0.66	**0.71**	0.65	**0.70**
General population	*Chi-square*[b]	13,228	2,515	13,314	7,761	2,716	4,373	2,282	2,189	4,087	5,523	5,216
sample (n = 1,280)	CFI	0.779	0.927	0.777	0.860	0.935	0.913	0.932	0.900	0.915	0.878	0.896
	TLI	0.767	0.919	0.767	0.852	0.929	0.908	0.925	0.889	0.908	0.868	0.889
	RMSEA	0.090	**0.077**	0.090	**0.080**	**0.073**	**0.069**	**0.067**	**0.068**	**0.073**	0.084	**0.071**
	SRMR	0.097	0.081	0.097	0.086	**0.078**	**0.074**	**0.072**	**0.075**	0.081	0.090	**0.079**
	% Item loadings >0.40	58.00%	**80.77%**	58.00%	72.09%	**89.29%**	**86.49%**	**89.29%**	78.57%	**82.35%**	71.43%	76.92%
	Mean coefficient alpha of factors	0.65	**0.70**	0.65	0.67	**0.75**	0.69	0.66	0.63	0.68	0.61	0.65

[a] These models included a hierarchical factor that subsumes four correlated factors.

[b] All chi-squared values were $P < 0.001$.

Table 3. Inter-factor Correlations Between the Factors Identified in the Russell-Smith et al. [2011] Three-Factor Model for the Undergraduate and General Population Samples

		Student sample		
		Social skill	Details/patterns	Communication/mindreading
General population sample	Social skill		0.108***	0.477***
	Details/patterns	0.122***		−0.055
	Comm./mindreading	0.447***	−0.077*	

*P < 0.05; ***P < 0.001.

the *Communication/Mindreading* and *Details/Patterns* dimensions were essentially uncorrelated. Correspondingly, based on the three-factor model described by Russell-Smith et al. [2011], we estimated the internal consistency of the total scale AQ scores at 0.39 and 0.37 for the undergraduate and general population samples, respectively, on the basis of the respective inter-subscale correlations, as recommended by Gignac [2014] for multidimensional scales.

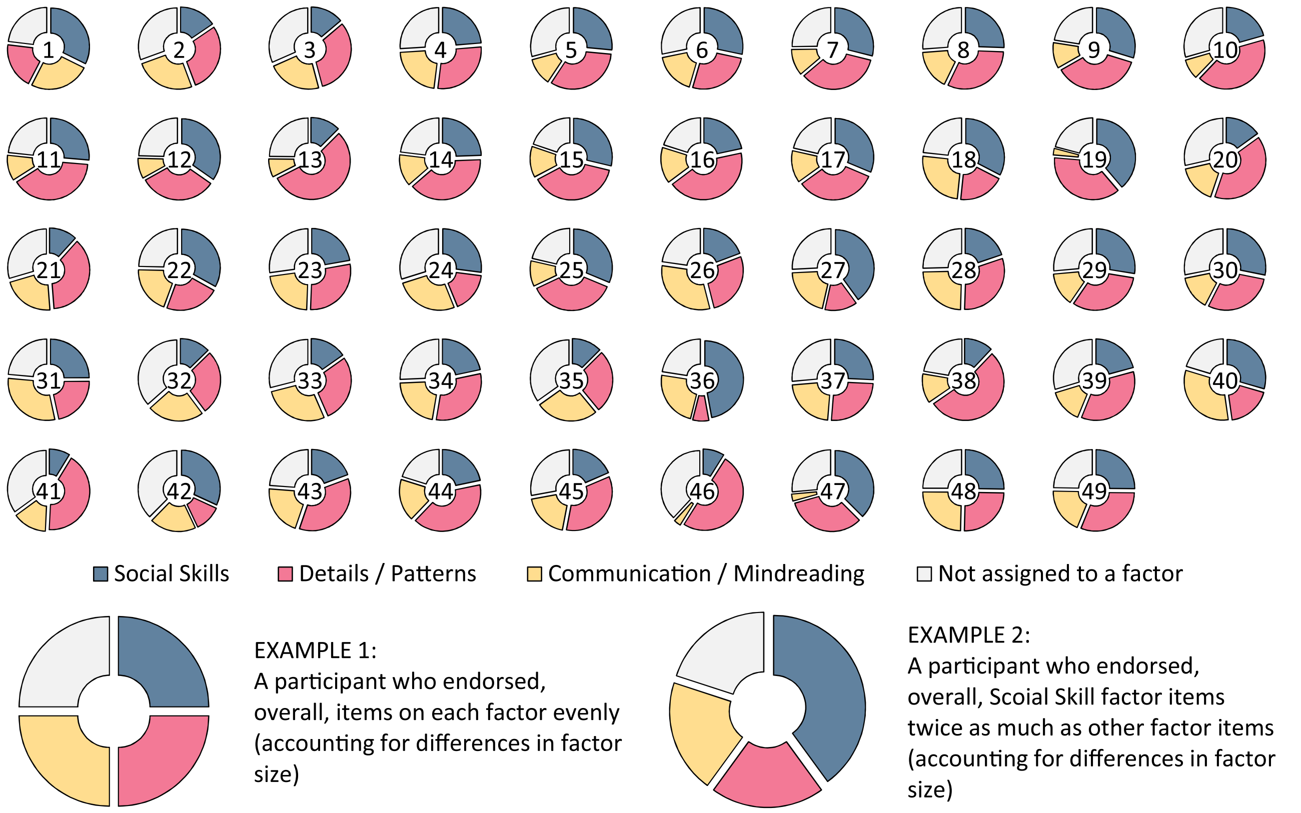
The variability in the associations between these dimensions suggests that the putative construct thought to underlie total scale AQ score is not as homogenous as originally believed, suggesting that substantial differences in the specific autistic trait dimension scores can be observed across people who have otherwise comparable total scale AQ scores. As outlined in the Methods section, we elected to visualize the homogeneity or heterogeneity of factor scores by representing the distribution of adjusted factor scores (i.e., adjusted to control for the different number of items on the factors) using pie-charts in which each chart represents a different participant, and each segment of the chart representing a different factor. We chose to examine participants from the undergraduate sample whose total scale AQ score matched the sample mean (M = 107, n = 49; detailed descriptive statistics can be found Table S3). An overview of the expression of individual factors for each participant, controlling for the number of items on each factor, is illustrated in Figure 2.

The visualization in Figure 2 clearly demonstrates substantial heterogeneity in the endorsement of items on each AQ factor, reflecting the weak inter-factor correlations that are present. For example, participants 26 and 37 show relatively similar levels of endorsement on each factor. Participants 19, 46, and 47, on the other hand, indicated that they had little difficulty with *Communication/Mindreading*, and thus their total scale AQ score primarily stemmed from endorsing the other factors. Finally, the adjusted factor scores for participants 13 and 46 were approximately 50% comprised of *Details/Patterns* items. As this factor has almost half the number of items as the *Social Skills* factor (7 vs. 13), these individuals' total scale AQ scores could be considered underestimated due to a combination of weak inter-factor correlations and unequal factor sizes.

Table 4. Series of Model Comparisons Assessing Measurement and Structural Invariance as a Function of Sample (Analysis 1), and Additionally as a Function of Sex (Analyses 2A and 2B) Using the Russell-Smith et al. [2011] Three-Factor Model

Model	χ^2	df	CFI	RMSEA (90% CI)	$\Delta\chi^2$	Δdf	ΔCFI	ΔRMSEA	Decision
1. Total Sample – Undergraduate vs. General Population									
Configural invariance	6,280	694	0.950	0.073 (0.072–0.075)	–	–	–	–	–
Metric (weak) invariance	6,373	747	0.950	0.071 (0.069–0.073)	92.71	53	0.000	−0.002	Accept
Scalar (strong) invariance	6,868	772	0.946	0.073 (0.071–0.074)	495.38	25	0.004	0.002	Accept
Residual (strict) invariance	7,075	800	0.944	0.073 (0.071–0.074)	207.18	28	0.002	0.000	Accept
Mean invariance	8,080	803	0.935	0.078 (0.076–0.080)	1,004.58	3	0.009	0.005	Prov. Accept
2A. Undergraduate Sample – Male vs. Female									
Configural invariance	3,856	694	0.959	0.073 (0.071–0.075)	–	–	–	–	Accept
Metric (weak) invariance	4,013	747	0.958	0.072 (0.070–0.074)	156.52	53	0.001	−0.001	Accept
Scalar (strong) invariance	4,263	772	0.955	0.073 (0.071–0.075)	249.91	25	0.003	0.001	Accept
Residual (strict) invariance	4,473	800	0.953	0.073 (0.071–0.076)	209.95	28	0.002	0.000	Accept
Mean invariance	4,684	803	0.950	0.075 (0.073–0.078)	211.27	3	0.003	0.002	Accept
2B. General Population Sample – Male vs. Female									
Configural invariance	3,152	694	0.934	0.074 (0.072–0.077)	–	–	–	–	Prov. Accept
Metric (weak) invariance	3,258	747	0.932	0.073 (0.070–0.075)	106.11	53	0.002	−0.001	Prov. Accept
Scalar (strong) invariance	3,571	772	0.925	0.075 (0.073–0.078)	312.72	25	0.007	0.002	Prov. Accept
Residual (strict) invariance	3,713	800	0.921	0.075 (0.073–0.078)	142.77	28	0.004	0.000	Prov. Accept
Mean invariance	3,922	803	0.916	0.078 (0.076–0.080)	208.48	3	0.005	0.003	Prov. Accept

Note. In addition to previously used fit indices [CFI > 0.95, RMSEA <0.08; Schweizer, 2010], ΔCFI >0.01 and ΔRMSEA >0.015 are also indicative of a violation of the invariance assumption [Chen, 2007; Cheung & Rensvold, 2002]. CFI, comparative fit index; RMSEA, root mean square error of approximation.



EXAMPLE 1:
A participant who endorsed, overall, items on each factor evenly (accounting for differences in factor size)

EXAMPLE 2:
A participant who endorsed, overall, Scoial Skill factor items twice as much as other factor items (accounting for differences in factor size)

Figure 2. A graphical visualization of the heterogeneity of AQ factor scores (defined using the Russell-Smith et al. [2011] three-factor model) within participants with identical Total Scale AQ scores. Each circle represents one of the 49 undergraduate participants with a Total Scale AQ score of 107. A hypothetical participant whose adjusted factor scores were identical would be illustrated like Example 1, whereas a hypothetical participant whose Social Skills adjusted factor score was twice the other adjusted factor scores would be illustrated like Example 2.

Discussion

Summary

As heterogeneity in the expression of autistic-like traits (and autism itself) is increasingly recognized and considered in research designs, the AQ is a potentially well-placed tool for quantifying the individual autistic-like trait dimensions seen in neurotypical individuals. However, the lack of consensus regarding the factor structure of the AQ is a potential barrier for achieving the levels of research quality and consistency that the field demands. In order to short-circuit the seemingly ongoing development and promotion of alternative AQ factor models, we conducted a comprehensive series of confirmatory factor analyses testing and comparing the mode fit of existing AQ factor models with the aim of highlighting those models with adequate psychometric properties. Critically, our series of analyses used methods not employed in the initial development of many of the models examined—namely, polychoric correlations (more appropriate for ordinal data) instead of Pearson correlations, large samples, and internal consistency calculated using inter-factor correlations. This last point was especially important, as nonpositive factor correlations undermine the interpretability of total scale AQ scores, and internal consistency calculated at the commonly used item level tends to overestimate the true internal consistency of a multidimensional scale.

Comparing AQ Factor Models

Similar to previous studies, we found a CFA on the original five-subscale model of the AQ [Baron-Cohen et al., 2001] yielded relatively poor fit indices. The model also showed a less-than-ideal average coefficient alpha (i.e., <0.70) across the five subscales. In addition, roughly 60% of items had a factor loading >0.40 across the two samples we examined which is well below the percentage for most of the alternative models tested. Based on these data, we suggest that the original five-subscale model is a psychometrically inadequate factor solution for the AQ. This has significant implications for how the inventory should be scored. The Hoekstra et al. [2008] model, which comprised the original model with the addition of four factors subsumed by a hierarchical factor, and the four-factor solution identified by Stewart and Austin [2009], also provided inadequate fits for the data in the current study. The remaining models all showed relatively "fair" levels of statistical close-fit across both samples. However, several of the fit indices for the five-factor models proposed by Lau, Gau, et al. [2013] and Lau, Kelly, and Peterson [2013] dipped below our cut offs for "fair" levels of fit for the general population sample [Schweizer, 2010].

The two models with the most psychometrically sound solutions were arguably the three-factor model proposed by Russell-Smith et al. [2011] and the five-factor model proposed by Kloosterman et al. [2011]. Across the fit indices, both showed comparably high levels of fit, with the three-factor model barely higher with respect to CFI and TLI, and the five-factor model barely lower with respect to RMSEA and SRMR across both samples. However, arguably the Russell-Smith et al. [2011] model may be favored, due to having a higher percentage of items (89.29% vs. 82.14%) with factor loadings >0.40 [the criterion for adequate factor loading used by Kloosterman et al., 2011] in the student sample, in addition to a higher mean coefficient alpha ($\alpha \approx 0.76$ vs. $\alpha \approx 0.67$) across each model's factors for both participant samples. The average coefficient alpha for Kloosterman et al.'s [2011] five-factor model was hindered mostly by the alpha for the *Restricted/Repetitive Behaviors* factor (0.47 in both samples). The authors themselves noted that the factor was "fragile" (p. 313) in their own analyses, with relatively poor internal consistency relative to the other factors.

Taken together, these data indicate that the three-factor model reported by Russell-Smith et al. [2011], that is, *Social Skill*, *Details/Patterns*, and *Communication/Mindreading*, should be considered by researchers and practitioners when scoring the AQ. However, a noticeable implication of this factor structure should become apparent at this point. Several other traits and behaviors associated with autism are notably absent. While many of the models included dimensions associated with autism, such as *Restricted Interests* or *Resistance to Change* (and to a lesser extent, *Imagination*), these factors generally showed much lower levels of internal consistency relative to *Social Skills*, *Attention to Detail*, and, to a lesser extent, the *Communication* factor ("A," "C," and "B" factors from Fig. 1, respectively). While this might suggest the domains themselves are not reliably present within the broader autism phenotype, it could also indicate that these factors require more/better items to achieve the stability seen in factors like *Social Skills*. Regardless, researchers interested in measuring these other trait dimensions are recommended to use alternative psychometric tests instead of the factor scores from poor-fitting AQ models.

Inter-Factor Correlations and Total Scale AQ Scores

Arguably, the most common use of the AQ scale is to obtain an overall AQ score that provides an indication of the general level of autistic-like traits expressed by an individual. However, our results suggest the interpretability of these scores may be questionable, given that the underlying factors do not all correlate positively. For example, with respect to the three-factor Russell-Smith et al. [2011] model, while the *Social Skill* and *Communication/Mindreading* factors had a moderate positive correlation, *Social Skill* showed only a weak positive correlation with the *Details/Patterns* factor, and the correlation between the *Details/Patterns* and *Communication/Mindreading* factors, depending on the sample, was either small and negative or statistically nonsignificant (summarized in Table 3). In fact, across all of the models, there was a mix of positive and negative correlations of varying strength and statistical significance (summarized in Table S2). While interpretation of the inter-factor correlations in the less well-fitting models should be interpreted with caution, it is noted that the negative or statistically nonsignificant correlation pairs commonly involved an *Attention to Detail* factor ("C" factor).

The variability across the inter-factor correlations suggests that certain autistic traits diverge from each other. While this is potentially useful as a means of identifying separate phenotypes within the broader autism phenotype, it calls into question the interpretability and utility of total scale AQ scores. This issue is especially apparent when considering the variability in item endorsement for each factor on the cross-section of undergraduate participants with a total scale AQ score of 107 in Figure 2. Specifically, while participants might be comparable at the total scale level, this ignores the vast heterogeneity with respect to participants' endorsement of the different trait dimensions. Thus, while high total scale AQ scores may be indicative of higher levels of autistic traits overall, it is not possible to say what these specific traits are based on the total scale AQ scores alone. To complicate matters further, if two factors are negatively correlated, higher scores on one factor are likely to be associated with *lower* scores on the other, which results in the "canceling out" of scores. Consequently, an individual who scores particularly high on a given factor can still have an otherwise moderate total scale AQ score, thus masking the elevated factor score. Overall, these issues are particularly problematic for researchers as reliance on total scale AQ scores inherently limits the interpretability of any experimental results.

The utility of total scale AQ scores is called into question further, on the basis of the estimation of the internal consistency reliability. Specifically, using the inter-factor correlations from the best-fitting model, that is, the Russell-Smith et al. [2011] three-factor model, we estimated omega hierarchical for the overall scale to be 0.39 and 0.37 for the student and general population samples, respectively. Such values are well below the 0.70 generally recommended for basic research for scale scores [Nunnally & Bernstein, 1994]. However, it would be misleading to suggest that the AQ is a unique multidimensional inventory with respect to its low total scale internal consistency reliability. In fact, based on a review of the omega hierarchical literature with personality type scales, Gignac and Kretzschmar [2017] suggested that omega hierarchical values of <0.20, 0.20–0.30, and >0.30 be considered relatively small, typical, and relatively large. Thus, the AQ's omega hierarchical of ≈0.40 would be considered relatively large, from this perspective. Importantly, however, both Gignac and Watkins [2013] and Reise, Bonifay, and Haviland [2013] recommended that omega hierarchical reach a minimum of 0.50, in absolute terms, for justifiable interpretations of the corresponding overall composite scores. Thus, the AQ total scores do not meet such a recommended absolute benchmark. Furthermore, a further complication with interpreting total AQ scores as representative of a global AQ trait is that two of the dimensions inter-correlate negatively (i.e., *Details/Patterns* and *Communication/Mindreading*), an arguably rare occurrence for multidimensional scales designed to measure a global trait.

The interpretation of our results is consistent with the growing acknowledgment that autism may be a heterogeneous condition, with possible subtypes and distinct clinical presentations [Masi, DeMayo, Glozier, & Guastella, 2017]. Thus, it may be appropriate to endorse a similar view with respect to autistic traits more broadly, especially considering the findings of nonpositively correlated AQ trait dimensions in the present study. Importantly, however, our discussion here should not be viewed as a specific criticism of previous research that has reported theoretically congruent results with the total AQ scores. Instead, we are proposing that more consistent, more insightful, and possibly more substantial findings may be achieved by analyzing AQ data at the subscale level, in comparison to relying upon analyses at the total scale level.

Comparisons to Clinical Diagnostic Tools

A primary reason driving research examining individual differences in autistic-like traits and behaviors is that the experimental findings of such studies are often qualitatively similar to results of studies that compare individuals with and without an autism diagnosis. However, while one might be tempted to extend the findings of the current study to clinical samples, it remains unclear if the three-factor structure of the AQ identified here for university undergraduates and the general population would be supported for clinically diagnosed autistic individuals. Unfortunately, it is not possible to draw on the findings of earlier work in this regard; while some previous factor analysis work on the AQ did incorporate large groups of autistic participants into their designs [Hoekstra et al., 2011; Lau, Kelly, & Peterson, 2013], these studies did not examine the extent to which the factor models they were proposing were valid for the clinical samples.

However, we believe it is important to investigate whether the nature of the relationships between specific autistic dimensions in clinical autism is comparable to those found in the samples of neurotypical individuals we investigated. Just as the presence of (essentially) uncorrelated or negatively correlated traits demonstrates the difficulty in using a unitary total scale score to meaningfully quantify autistic traits in neurotypical individuals, it is possible that similar associations between symptom dimensions in clinical autism complicate the use of screener instruments and the diagnostic process. Commonly used diagnostic tools, such as the Autism Diagnostic Observation Schedule [ADOS-2; Lord, Rutter, DiLavore, & Risi, 2003] and Autism Diagnostic Interview-Revised [ADI-R; Lord, Rutter, & Le Couteur, 1994], assess for distinct symptom domains to establish reliable diagnoses, and an extensive review of factor analyses of autistic symptoms as measured by such diagnostic tools confirms that at least two separate dimensions are observable in autistic individuals—a social/communication factor and a repetitive and restrictive behaviors and interests factor [for a review, see Shuster, Perry, Bebko, & Toplak, 2014].

It is this consideration of individual dimensions that separates how diagnostic tools, such as the ADOS-2 and ADI-R, and the AQ are used. With respect to the use of diagnostic tools, individuals must exceed cutoff criteria for multiple dimensions (symptom clusters) before a diagnosis of ASD can be made. An individual who meets cutoffs for certain symptom dimensions, but not others, may not receive an ASD diagnosis, but might receive an alternative diagnosis [e.g., pragmatic language disorder; Whitehouse, Evans, Eapen, & Wray, 2018]. In contrast, the AQ is generally treated as a single, unitary dimension in differentiating individuals. In this sense, autistic participants and neurotypical, "high autistic trait" participants are not necessarily comparable, as the neurotypical participants may not demonstrate the full range of autistic dimensions. Perhaps when using the AQ for research purposes, it would be more appropriate to ensure that participants exceed certain thresholds across all three factors identified in the Russell-Smith et al. [2011] factor model before applying the label of "High AQ" to participants. Otherwise, it may be improper to suggest that a neurotypical, "high autistic trait" group of participants is qualitatively comparable to an autistic group that has demonstrated autistic behaviors across all of the key dimensions assessed by instruments such as the ADOS-2 and ADI-R.

Conclusions

There is substantial variability in the overall quality of the numerous factor structures for the AQ that have been proposed to date. We note that the original five-subscale model [Baron-Cohen et al., 2001] is not statistically supported, as others have previously reported, but we also highlight several other models that are also of questionable utility. While we also identified several models that show relatively fair levels of model fit in our large samples, the results of our analyses recommend the use of the Russell-Smith et al. [2011] three-factor structure given it demonstrates adequate fit indices across both samples and does not contain any particularly "weak" factors—a common finding across the models that were numerous in terms of factors. Furthermore, multigroup factorial invariance analysis across the undergraduate and general population samples indicated that this factor structure differs little between the two samples and is therefore appropriate for use in both population types.

The present study also highlights several issues with certain aspects of the scale more broadly. Namely, there is strong evidence for divergence between the factors of the AQ across all of the models we identified in the literature. This calls into question the interpretability of total scale AQ scores as they may be derived from uncorrelated or even negatively correlated underlying factors. Finally, using a more appropriate method of calculating internal consistency given the divergence of factors within the AQ [Gignac, 2014], we estimate coefficient omega hierarchical to be roughly ≈ 0.38, which is far lower than the minimum of 0.50 required for justifiable interpretations of the corresponding overall composite scores [Gignac & Watkins, 2013; Reise et al., 2013].

Based on the findings of the present study, we cannot advise researchers wishing to measure autistic traits in neurotypical individuals to use the AQ in its original five-subscale form. However, factor scores calculated using the factors identified in certain models, such as the three-factor Russell-Smith et al. [2011] model, may be retained and used for research purposes. It is also clear from the various models identified in the literature that the AQ taps into several different trait domains. The lack of consistency with respect to the extraction of several of these factors highlights the need for the development of alternative measures that can better encapsulate these different dimensions.

Acknowledgments

We are grateful to the Raine Study participants and their families and we thank the Raine Study Team for study coordination and data collection. The core management of the Raine Study is funded by the National Health and Medical Research Council, the University of Western Australia, Curtin University, Telethon Kids Institute, Women and Infants Research Foundation, Edith Cowan University, Murdoch University, The University of Notre Dame Australia, and the Raine Medical Research Foundation. This research was also supported by an Australian Research Council Discovery Grant (DP190103286) to M.T.M., A.J.O.W., T.A.W.V., and G.E.G.

Conflict of Interest

The authors declare that there were no identifiable conflict of interests with respect to the authorship and publication of this article.

Author Contributions

M.C.W.E., M.T.M., and G.E.G. jointly developed the study concept and design. Data analyses were performed by M.C.W.E. with assistance from G.E.G. The manuscript was drafted by M.C.W.E and critical revisions were provided by G.E.G., T.A.W.V., A.J.O.W. and M.T.M. All authors approved the final version of the manuscript prior to submission.

References

Abu-Akel, A., Allison, C., Baron-Cohen, S., & Heinke, D. (2019). The distribution of autistic traits across the autism spectrum: Evidence for discontinuous dimensional subpopulations underlying the autism continuum. Molecular Autism, 10, 1–13. https://doi.org/10.1186/s13229-019-0275-3

Asún, R. A., Rdz-Navarro, K., & Alvarado, J. M. (2016). Developing multidimensional Likert scales using item factor analysis: The case of four-point items. Sociological Methods and Research, 45 (1), 109–133. https://doi.org/10.1177/0049124114566716

Austin, E. J. (2005). Personality correlates of the broader autism phenotype as assessed by the Autism Spectrum Quotient (AQ). Personality and Individual Differences, 38(2), 451–460. https://doi.org/10.1016/j.paid.2004.04.022

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The autism-spectrum quotient (AQ): Evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. Journal of Autism and Developmental Disorders, 31(1), 5–17. https://doi.org/10.1023/A:1005653411471

Bishop, D. V. M., Maybery, M., Wong, D., Maley, A., & Hallmayer, J. (2006). Characteristics of the broader phenotype in autism: A study of siblings using the children's communication checklist-2. American Journal of Medical Genetics - Neuropsychiatric Genetics, 141 B(2), 117–122. https://doi.org/10.1002/ajmg.b.30267

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. Structural Equation Modeling, 14 (3), 464–504. https://doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. Structural Equation Modeling, 9(2), 233–255. https://doi.org/10.1207/S15328007SEM0902

Constantino, J. N., & Gruber, C. P. (2012). Social responsiveness scale–second edition (SRS-2). Torrance, CA: Western Psychological Services.

Constantino, J. N., & Todd, R. D. (2003). Autistic traits in the general population: A twin study. Archives of General Psychiatry, 60, 524–530. https://doi.org/10.1001/archpsyc.60.5.524

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. Practical Assessment, Research & Education, 10, 1–9. https://doi.org/10.1.1.110.9154

Davis, J., McKone, E., Zirnsak, M., Moore, T., O'Kearney, R., Apthorp, D., & Palermo, R. (2017). Social and attention-to-detail subclusters of autistic traits differentially predict looking at eyes and face identity recognition ability. British Journal of Psychology, 108(1), 191–219. https://doi.org/10.1111/bjop.12188

DiStefano, C. (2002). A multilevel structural equation model for dyadic data. Structural Equation Modeling, 9(3), 327–346. https://doi.org/10.1207/S15328007SEM0903

Feczko, E., Balba, N. M., Miranda-Dominguez, O., Cordova, M., Karalunas, S. L., Irwin, L., … Fair, D. A. (2018). Subtyping cognitive profiles in Autism Spectrum Disorder using a Functional Random Forest algorithm. NeuroImage, 172, 674–688. https://doi.org/10.1016/j.neuroimage.2017.12.044

Fountain, C., Winter, A. S., & Bearman, P. S. (2012). Six developmental trajectories characterize children with autism. Pediatrics, 129(5), e1112–e1120. https://doi.org/10.1542/peds.2011-1601

Frazier, T. W., Ratliff, K. R., Gruber, C., Zhang, Y., Law, P. A., & Constantino, J. N. (2014). Confirmatory factor analytic structure and measurement invariance of quantitative autistic traits measured by the Social Responsiveness Scale-2. Autism, 18(1), 31–44. https://doi.org/10.1177/1362361313500382

Freeth, M., Sheppard, E., Ramachandran, R., & Milne, E. (2013). A cross-cultural comparison of autistic traits in the UK, India and Malaysia. Journal of Autism and Developmental Disorders, 43 (11), 2569–2583. https://doi.org/10.1007/s10803-013-1808-9

Geschwind, D. H., & State, M. W. (2015). Gene hunting in autism spectrum disorder: On the path to precision medicine. The Lancet Neurology, 14(11), 1109–1120. https://doi.org/10.1016/S1474-4422(15)00044-7

Gignac, G. E. (2014). On the inappropriateness of using items to calculate total scale score reliability via coefficient alpha for multidimensional scales. European Journal of Psychological Assessment, 30(2), 130–139. https://doi.org/10.1027/1015-5759/a000181

Gignac, G. E., & Kretzschmar, A. (2017). Evaluating dimensional distinctness with correlated-factor models: Limitations and suggestions. Intelligence, 62, 138–147. https://doi.org/10.1016/j.intell.2017.04.001

Gignac, G. E., & Watkins, M. W. (2013). Bifactor modeling and the estimation of model-based reliability in the WAIS-IV. Multivariate Behavioral Research, 48(5), 639–662. https://doi.org/10.1080/00273171.2013.804398

Green, S. B., & Yang, Y. (2015). Evaluation of dimensionality in the assessment of internal consistency reliability: Coefficient alpha and omega coefficients. Educational Measurement: Issues and Practice, 34(4), 14–20. https://doi.org/10.1111/emip.12100

Hanel, P. H. P., & Vione, K. C. (2016). Do student samples provide an accurate estimate of the general public? PLoS One, 11 (12), 1–10. https://doi.org/10.1371/journal.pone.0168354

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. Applied Psychological Measurement, 9(2), 139–164. https://doi.org/10.1177/014662168500900204

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? Behavioral and Brain Sciences, 33(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Hirschfeld, G., von Brachel, R., & Thielsch, M. (2014). Selecting items for Big Five questionnaires: At what sample size do factor loadings stabilize? Journal of Research in Personality, 53, 54–63. https://doi.org/10.1016/j.jrp.2014.08.003

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(8), 1555–1566. https://doi.org/10.1007/s10803-008-0538-x

Hoekstra, R. A., Vinkhuyzen, A. A. E., 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(5), 589–596. https://doi.org/10.1007/s10803-010-1073-0

Hogarty, K. Y., Hines, C. V., Kromrey, J. D., Ferron, J. M., & Mumford, K. R. (2005). The quality of factor solutions in exploratory factor analysis: The influence of sample size, communality, and overdetermination. Educational and Psychological Measurement, 65(2), 202–226. https://doi.org/10.1177/0013164404267287

Holgado-Tello, F. P., Chacón-Moscoso, S., Barbero-García, I., & Vila-Abad, E. (2010). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. Quality & Quantity, 44(1), 153–166. https://doi.org/10.1007/s11135-008-9190-y

Hurst, R. M., Mitchell, J. T., Kimbrel, N. A., Kwapil, T. K., & Nelson-Gray, R. O. (2007). Examination of the reliability and factor structure of the Autism Spectrum Quotient (AQ) in a non-clinical sample. Personality and Individual Differences, 43(7), 1938–1949. https://doi.org/10.1016/j.paid.2007.06.012

Kim, S. H., Macari, S., Koller, J., & Chawarska, K. (2016). Examining the phenotypic heterogeneity of early autism spectrum disorder: Subtypes and short-term outcomes. Journal of Child Psychology and Psychiatry and Allied Disciplines, 57(1), 93–102. https://doi.org/10.1111/jcpp.12448

Kloosterman, P. H., Keefer, K. V., Kelley, E. A., Summerfeldt, L. J., & Parker, J. D. A. (2011). Evaluation of the factor structure of the Autism-Spectrum Quotient. Personality and Individual Differences, 50(2), 310–314. https://doi.org/10.1016/j.paid.2010.10.015

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. International Journal of Psychological Research, 3(1), 111. https://doi.org/10.21500/20112084.857

Landry, O., & Chouinard, P. A. (2016). Why we should study the broader autism phenotype in typically developing populations. Journal of Cognition and Development, 17(4), 584–595. https://doi.org/10.1080/15248372.2016.1200046

Lau, W. Y. P., Gau, S. S. F., Chiu, Y. N., Wu, Y. Y., Chou, W. J., Liu, S. K., & Chou, M. C. (2013). Psychometric properties of the Chinese version of the Autism Spectrum Quotient (AQ). Research in Developmental Disabilities, 34(1), 294–305. https://doi.org/10.1016/j.ridd.2012.08.005

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 (12), 2807–2815. https://doi.org/10.1007/s10803-013-1827-6

Lord, C., Bishop, S., & Anderson, D. (2015). Developmental trajectories as autism phenotypes. American Journal of Medical Genetics, Part C: Seminars in Medical Genetics, 169(2), 198–208. https://doi.org/10.1002/ajmg.c.31440

Lord, C., Rutter, M., DiLavore, P. C., & Risi, S. (2003). Autism diagnostic observation schedule: ADOS. CA: Western Psychological Services Los Angeles.

Lord, C., Rutter, M., & Le Couteur, A. (1994). Autism Diagnostic Interview-Revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. Journal of Autism and Developmental Disorders, 24(5), 659–685.

Mandy, W. P. L., & Skuse, D. H. (2008). Research Review: What is the association between the social-communication element of autism and repetitive interests, behaviours and activities? Journal of Child Psychology and Psychiatry and Allied Disciplines, 49(8), 795–808. https://doi.org/10.1111/j.1469-7610.2008.01911.x

Masi, A., DeMayo, M. M., Glozier, N., & Guastella, A. J. (2017). An overview of autism spectrum disorder, heterogeneity and treatment options. Neuroscience Bulletin, 33(2), 183–193. https://doi.org/10.1007/s12264-017-0100-y

Mellenbergh, G. J. (1989). Item bias and item response theory. International Journal of Educational Research, 13(2), 127–143. https://doi.org/10.1016/0883-0355(89)90002-5

Nunnally, J. C., & Bernstein, I. H. (1994). In I. H. Bernstein (Ed.), Psychometric theory (3rd ed.). New York: McGraw-Hill.

Peterson, R. A., & Merunka, D. R. (2014). Convenience samples of college students and research reproducibility. Journal of Business Research, 67(5), 1035–1041. https://doi.org/10.1016/j.jbusres.2013.08.010

Pickles, A., Anderson, D. K., & Lord, C. (2014). Heterogeneity and plasticity in the development of language: A 17-year follow-up of children referred early for possible autism. Journal of Child Psychology and Psychiatry and Allied Disciplines, 55(12), 1354–1362. https://doi.org/10.1111/jcpp.12269

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. Journal of Personality Assessment, 95(2), 129–140. https://doi.org/10.1080/00223891.2012.725437

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. Psychological Assessment, 12(3), 287–297. https://doi.org/10.1037/1040-3590.12.3.287

Ronald, A., Happé, F., Bolton, P., Butcher, L. M., Price, T. S., Wheelwright, S., … Plomin, R. (2006). Genetic heterogeneity between the three components of the autism spectrum: A twin study. Journal of the American Academy of Child and Adolescent Psychiatry, 45(6), 691–699. https://doi.org/10.1097/01.chi.0000215325.13058.9d

Ronald, A., & Hoekstra, R. A. (2011). Autism spectrum disorders and autistic traits: A decade of new twin studies. American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics, 156(3), 255–274. https://doi.org/10.1002/ajmg.b.31159

Russell-Smith, S. N., Maybery, M. T., & Bayliss, D. M. (2011). Relationships between autistic-like and schizotypy traits: An analysis using the Autism Spectrum Quotient and Oxford-Liverpool Inventory of Feelings and Experiences. Personality and Individual Differences, 51(2), 128–132. https://doi.org/10.1016/j.paid.2011.03.027

Russell-Smith, S. N., Maybery, M. T., Bayliss, D. M., & Sng, A. A. H. (2012). Support fora link between the local processing bias and social deficits in Autism: An investigation of embedded figures test performance in non-clinical individuals. Journal of Autism and Developmental Disorders, 42(11), 2420–2430. https://doi.org/10.1007/s10803-012-1506-z

Schweizer, K. (2010). Some guidelines concerning the modeling of traits and abilities in test construction. European Journal of Psychological Assessment, 26(1), 1–2. https://doi.org/10.1027/1015-5759/a000001

Shuster, J., Perry, A., Bebko, J., & Toplak, M. E. (2014). Review of factor analytic studies examining symptoms of autism spectrum disorders. Journal of Autism and Developmental Disorders, 44(1), 90–110. https://doi.org/10.1007/s10803-013-1854-3

Skuse, D. H., Mandy, W. P. L., & Scourfield, J. (2005). Measuring autistic traits: Heritability, reliability and validity of the Social and Communication Disorders Checklist. British Journal of Psychiatry, 187(6), 568–572. https://doi.org/10.1192/bjp.187.6.568

Stevens, J. P. (2009). Applied multivariate statistics for the social sciences (5th ed.). New York: Routledge.

Stevenson, J. L., & Hart, K. R. (2017). Psychometric properties of the Autism-Spectrum Quotient for assessing low and high levels of autistic traits in college students. Journal of Autism and Developmental Disorders, 47(6), 1838–1853. https://doi.org/10.1007/s10803-017-3109-1

Stewart, M. E., & Austin, E. J. (2009). The structure of the Autism-Spectrum Quotient (AQ): Evidence from a student sample in Scotland. Personality and Individual Differences, 47(3), 224–228. https://doi.org/10.1016/j.paid.2009.03.004

Straker, L., Mountain, J., Jacques, A., White, S., Smith, A., Landau, L., … Eastwood, P. (2017). Cohort profile: The western australian pregnancy cohort (Raine) study–generation 2. International Journal of Epidemiology, 46(5), 1384–1385j. https://doi.org/10.1093/ije/dyw308

Taylor, L. J., Maybery, M. T., Grayndler, L., & Whitehouse, A. J. O. (2014). Evidence for distinct cognitive profiles in autism spectrum disorders and specific language impairment. Journal of Autism and Developmental Disorders, 44(1), 19–30. https://doi.org/10.1007/s10803-013-1847-2

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. Organizational Research Methods, 3(1), 4–70. https://doi.org/10.1177/109442810031002

Whitehouse, A., Evans, K., Eapen, V., & Wray, J. (2018). A national guideline for the assessment and diagnosis of autism spectrum disorder in Australia. Brisbane: Cooperative Research Centre for Living with Autism.

Whitten, A., Unruh, K. E., Shafer, R. L., & Bodfish, J. W. (2018). Subgrouping autism based on symptom severity leads to differences in the degree of convergence between core feature domains. Journal of Autism and Developmental Disorders, 48(6), 1908–1919. https://doi.org/10.1007/s10803-017-3451-3

Wu, A. D., Li, Z., & Zumbo, B. D. (2002). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with

TIMSS data - practical assessment, research & evaluation. Practical Assessment, Research & Evaluation, 12(3), 1–26.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table S1. Standardized coefficient alpha for factors identified in each model separated by sample.

Table S2. Inter-factor correlations for each model and cohort.

Table S3. Descriptive statistics for three-factor Russell-Smith et al. [2011] factor scores and adjusted factor scores (in parentheses) for undergraduate participants with a Total Scale AQ score of 107 (*n* = 49).