Journal of Autism and Developmental Disorders (2019) 49:1949–1965
https://doi.org/10.1007/s10803-019-03880-x

ORIGINALPAPER

Psychometric Assessments of Three Self-Report Autism Scales (*AQ*, *RBQ-2A*, and *SQ*) for General Adult Populations

Ronnie Jia[1] · Zachary R. Steelman[2] · Heather H. Jia[1]

Published online: 21 January 2019

Abstract

This study assesses the psychometric properties of three self-report measures of autistic-like tendencies in the general adult population: *autistic spectrum quotient (AQ), adult repetitive behaviours questionnaire-2 (RBQ-2A)*, and *systemizing quotient (SQ)*. Three rounds of development and testing using different U.S. and global samples led to three instruments that are psychometrically sound, parsimonious, and generalizable across populations. The resulting *AQ-9*, consisting of two factors: *social communication* and *attention to detail*, now mirrors the current dual diagnostic criteria in the *DSM-5*. The *RBQ-2A-R* has now been refined through CFA for the first time. The new *SQ-7* scale also has updated content. All three refined scales demonstrate satisfactory psychometric validity and parsimony and now provide evidence of their appropriateness for empirical research.

Keywords Autism · AQ · RBQ-2A · SQ · Mechanical Turk · Factorial validity

Introduction

While diagnosticians make individual diagnoses of the autism spectrum condition (ASC) in clinical settings based on direct observations of the patients and reports by caregivers, empirical researchers seeking to gauge autistic-like tendencies in the nonclinical adult population often rely on self-report measures. Our literature search has identified three such scales: *autistic spectrum quotient* (*AQ*; Baron-Cohen et al. 2001), *adult repetitive behaviors questionnaire-2* (*RBQ-2A*; Barrett et al. 2015), and *systemizing quotient* (*SQ*; Baron-Cohen et al. 2003).[1] These self-report measures are not intended to be diagnostic instruments, but have been used as screening tools and administered along with other scales in empirical survey-based research.

Among these, the most widely used scale is likely the *AQ* (Baron-Cohen et al. 2001), a comprehensive measure of autism-like symptoms. Since its publication, the article containing the original scale has been cited thousands of times, and its influence continues to grow (Table 1). However, as elaborated in later sections, the factor structure of the *AQ* remains inconclusive despite its popularity and frequent use in prior literature. Further, neither the original scale nor its abbreviated versions are in keeping with the current two-factor ASC diagnostic criteria specified in the *diagnostic and statistical manual of mental disorders* (*DSM-5*; American Psychiatric Association 2013), potentially leading to a disconnect between academic research on ASC and clinical practice.

Complementing the *AQ*, the *RBQ-2A* (Barrett et al. 2015) assesses a specific set of autistic symptoms, i.e., restricted

Dr. Ronnie Jia is an Associate Professor at Illinois State University, Normal, Illinois, USA.
Dr. Zachary R. Steelman is an Assistant Professor at University of Arkansas, Fayetteville, Arkansas, USA.
Dr. Heather H. Jia is an Associate Professor at Illinois State University, Normal, Illinois, USA.

✉ Ronnie Jia
RonnieJia@gmail.com

Zachary R. Steelman
ZSteelman@walton.uark.edu

Heather H. Jia
HHJia@ilstu.edu

[1] Illinois State University, 304 Old Union, Normal, IL 61761, USA

[2] University of Arkansas, Fayetteville, AR 72701, USA

[1] There exists another adult, self-report scale, the *Autism Spectrum Disorder in Adults Screening Questionnaire* (ASDASQ; Nylander and Gillberg 2001). It was not included in this study because it "used a Scandinavian definition" of autism, rather than *DSM* or *ICD-10*, which "limits the evidence for its value" (Carpenter 2012, p. 123) and relationship to prior literature on autism-like symptoms.

Table 1 An overview of three adult, self-report autism-related measures

Scale	Origin	Citations in Google Scholar (as of January 2, 2019)	
		Total	Since 2017
Autistic spectrum quotient (*AQ*)	Baron-Cohen et al. (2001)	3425	934
Systemizing quotient (*SQ*)	Baron-Cohen et al. (2003)	808	131
Adult repetitive behaviors questionnaire-2 (*RBQ-2A*)	Barrett et al. (2015)	16	14

and repetitive behaviors. Individuals with high autistic tendencies may have limited insight into their social and communication challenges,[2] which could bias their responses in the non-autism direction (Bishop and Seltzer 2012). Therefore, by focusing on one of the most directly observable aspects of autism, this scale does not rely on the respondents' introspection and is less likely influenced by their limited insight (Lewis and Bodfish 1998). Given its relative newness, we are not aware of any follow-up research that has retested the psychometric properties of the *RBQ-2A* using confirmatory factor analysis (CFA).

The *SQ* (Baron-Cohen et al. 2003), developed based on the Empathizing-Systemizing Theory (Baron-Cohen 2002), has also been frequently used over the years (Table 1). However, similar to the *AQ*, the *SQ* also has factorial validity issues in addition to some of its items appearing outdated and needing to be revised or removed to remain current (e.g., "I find it difficult to learn how to program video recorders").

In this research, we examine the *AQ, RBQ-2A*, and *SQ* for their appropriateness for empirical survey-based research using two key criteria: psychometric validity and parsimony.

To be appropriate for empirical survey-based research, a measurement instrument must possess satisfactory *psychometric validity*, such as reliability and factorial validity (Hair et al. 1998). While most researchers follow the same guidelines on scale reliability (e.g., Cronbach's $\alpha \geq 0.70$, Nunnally 1978), there has been less attention and consistency in the types of evidence required to demonstrate factorial validity (e.g., convergent validity, discriminant validity, measurement invariance), especially when evaluating these autism-related scales.

Many studies have relied on exploratory techniques, such as exploratory factor analysis (EFA) and principle component analysis (PCA), which often lead to significant overfactoring (Frazier and Youngstrom 2007), rather than CFA, which can evaluate alternative *a priori* factor structures for the best model fit and effectively establish factorial validity (e.g., MacCallum et al. 1992). In addition to overall model fit, items should also have high loadings on the intended factors (convergent validity) and low loadings on other factors (discriminant validity). Thresholds for individual item loadings have been recommended (e.g., 0.71 = excellent, 0.63 = very good, 0.55 = good, 0.45 = fair, and 0.32 = poor; Tabachnick and Fidell 2007). While it is critical for a factor to have at least three items (Kline 2010; Velicer and Fava 1998), it is desirable for a factor to have at least four loadings of 0.60 or higher (Guadagnoli and Velicer 1988).

Besides psychometric validity, another characteristic of a desirable measurement scale for empirical research, especially in survey-based research, is *parsimony*. In survey studies that include a large number of measurement items, researchers may be rightfully concerned about participants' response burden, which can lead to decreased data quality and response rates (Veale and Williams 2015). While researchers need valid measures with strong loadings and domain coverage of their focal constructs, scale parsimony is also necessary to encourage respondent participation and reduce response burden and bias (e.g., Deutskens et al. 2004; Dillman 2000).

In sum, while the above three scales have been used repeatedly in research, they have presented either factorial validity issues (*AQ* and *SQ*), need updating for a modern context (*SQ*), or have never been further tested using CFA (*RBQ-2A*). Given their influence in the literature, it is important to ensure their validity and refine them as needed to provide rigorous and consistent instruments for future empirical research. In this work, we refine these three scales using samples from nonclinical, general adult populations, such as U.S. college students and members of an online crowdsourcing platform (MTurk). The resulting measures are not intended to be diagnostic instruments for ASC, but as measurement scales in empirical survey-based studies examining the relationships between autism-like symptoms and other constructs.

In the following sections, we first review the previous scale development efforts related to these three autism-related measures and discuss their limitations based on the criteria of psychometric validity and parsimony. Then,

[2] The reverse may also be true as research on the "Double Empathy Problem" of autism has shown that neurotypical individuals may also struggle to read the emotions of autistic participants (e.g., Milton 2012). Thus, the issue of limited insight is arguably a mutual one.

we iteratively test and refine these scales for their robustness and consistency through three consecutive studies: we examine the psychometric properties of the existing scales in Study 1 with a sample of U.S. college students, further refine them through adding or removing items in Study 2 using a worldwide sample of MTurk members, and conduct a final psychometric test of the revised scales in Study 3 with a sample of U.S.-based MTurk members. In these consecutive evaluations, we assess scale reliability, convergent and discriminant validity, factorial invariance, and nomological validity in terms of their intercorrelations with relevant Big Five personality traits that have been established in prior autism research. Evidence for factorial invariance will also enable us to conduct cross-group comparisons, such as gender-based differences between male and female participants, which have also been widely reported in prior autism research. This collective set of tests will allow us to provide a psychometrically valid, parsimonious, and consistent measure in line with prior academic research and clinical practice. We conclude with a discussion of the benefits and usage of the final scales, their limitations, and recommendations for future research.

Prior Scale Development Efforts

Autistic Spectrum Quotient (AQ)

Likely the most frequently used autism-related self-report measure, the original *AQ* scale (Baron-Cohen et al. 2001) consists of 50 items in five factors: *social skill, attention switching, attention to detail, communication*, and *imagination*. However, despite its popularity, the 5-factor structure has received limited support in subsequent research, where a 3-factor model has been identified and replicated (e.g., Austin 2005; Hurst et al. 2007; Kloosterman et al. 2011).

In the first EFA of the *AQ-50*, Austin (2005) extracted a 26-item solution (*AQ-26*) in 3 factors, namely *social skills* ($\alpha = 0.85$), *details/patterns* ($\alpha = 0.70$) and *communication/mindreading* ($\alpha = 0.66$). Using PCA, Hurst et al. (2007) largely replicated Austin's (2005) 3-factor solution, however, the third factor, *communication/mindreading*, exhibited low internal consistency ($\alpha = 0.42$) as well as low item loadings across its four items (0.33, 0.53, 0.58, and 0.61). Though Hurst et al. (2007) noted that an alternative 2-factor solution could have been supported, the third factor was nonetheless retained due to their desire to "link the three identified factors to the autism triad, consistent with the [then] current diagnostic criteria" (p. 1947).

In a CFA test of the *AQ-50*, Kloosterman et al. (2011) also found support for a 3-factor model, including *social skills, attention to detail*, and *communication/mindreading*. However, similar to Austin (2005) and Hurst et al. (2007), the third factor exhibited lower reliability ($\alpha = 0.65$), and only two of its five items had loadings over 0.50.

This set of studies, all reporting a three-factor solution with internal consistency and/or convergent validity in the *communication/mindreading* factor lower than the recommended thresholds (e.g., Guadagnoli and Velicer 1988; Tabachnick and Fidell 2007), provides cumulative evidence in support of a 2-factor *AQ* model (*social skills* and *attention to detail*), which, echoing Hurst et al.'s (2007) call, would be consistent with the current two-factor ASC diagnostic criteria specified in *DSM-5*: "persistent deficits in social communication and social interaction" and "restricted, repetitive patterns of behavior, interests, or activities" (American Psychiatric Association 2013). Therefore, continued evaluation and refinement of the *AQ* is needed in keeping with the current *DSM-5* criteria, which will allow for the further alignment between academic research and clinical practice.

Another noteworthy effort to achieve a parsimonious *AQ* measure is Allison et al.'s (2012) *AQ-10*, which was proposed as a rapid screening tool and was constructed by choosing the two items with the highest discrimination index values from each of the five *AQ-50* scales (Baron-Cohen et al. 2001). However, evidence for its factorial validity was not reported, and its 5-factor structure is unlikely to hold in view of findings in Austin (2005), Hurst et al. (2007), and Kloosterman et al. (2011). Additionally, since each of the subscales consist of only two items despite recommendations for at least three items per factor (Kline 2010; Velicer and Fava 1998), they may suffer from low reliability as well as model estimation problems (Kline 2010), which can hinder their value when used in survey-based research.

Since the factorial validity of the *AQ-10* has never been reported in the literature, and the scale is frequently used in research (e.g., Jackson et al. 2018), it is necessary that its psychometric soundness be empirically tested. Despite its potential shortcomings, it still has the distinct advantage of being the briefest *AQ* measure. In view of its parsimony, it was still used as a starting point for constructing a parsimonious and psychometrically sound *AQ* measure that is also theoretically linked to the *DSM-5* diagnostic criteria. (The *AQ-50* was not chosen because it is neither parsimonious nor psychometrically sound.)

Adult Repetitive Behavior Questionnaire (RBQ-2A)

The *RBQ-2A* (Barrett et al. 2015) focuses on a specific set of autism-related behaviors, i.e., restricted and repetitive behaviors. Developed using PCA, it consists of 14 items in two subscales: *repetitive motor behavior* (RMB) and *insistence on sameness* (IoS). RMB includes motor mannerisms, sensory seeking behaviors, and repetitive use of objects, while IoS is characterized by compulsions, rituals, and difficulties with changes in routine (Cuccaro et al. 2003).

Table 2 Road map of multi-study scale refinement and validation

Study	Study 1—initial evaluation ($N_1 = 207$)	Study 2—further refinement ($N_2 = 355$)	Study 3—final validation ($N_3 = 442$)
Scale administered	AQ-10 (original)	AQ-10 (refined)	AQ-9 (refined)
	SQ-8 (original)	SQ-8 (refined)	SQ-7 (refined)
	RMB-2A (original)	RMB-2A (original)	RMB-2A-R (refined)
			Big Five—extroversion (original)
			Big Five—neuroticism (original)
Goal	To establish a baseline by testing the briefest existing versions of these measures to identify/confirm psychometric issues	To refine scales by adding, dropping, and revising items to establish evidence of satisfactory psychometrics and parsimony	To retest and validate the refined scales through evidence of satisfactory psychometrics, parsimony, measurement invariance, and nomological validity
Analysis	Reliability	CFA	CFA
	EFA (for AQ-10)	Convergent validity	Convergent validity
	CFA (for SQ-8 and RMB-2A)	Discriminant validity	Discriminant validity
	Convergent validity	Goodness of fit	Goodness of fit
	Discriminant validity	Reliability	Measurement invariance
	Goodness of fit		Reliability
			Descriptive statistics
			Nomological validity
			Correlations with big five personality traits
			Known gender differences

However, there is limited evidence of its factorial validity as three items have "poor" loadings ($\lambda < 0.45$, Tabachnick and Fidell 2007) and another three item loadings are in the "fair" range ($\lambda < 0.55$, Tabachnick and Fidell 2007). Unfortunately, no follow-up research using this scale has retested and reported its psychometric properties using CFA. It is therefore necessary to further validate and refine the scale for future empirical research.

Systemizing Quotient (SQ)

The 40-item *SQ* scale (*SQ-40*, Baron-Cohen et al. 2003) was proposed as a measure of autistic tendencies based on the Empathizing-Systemizing Theory (Baron-Cohen 2002), which posits that individuals with high autistic tendencies have impaired empathizing, but superior systemizing, which refers to the drive to analyze, control, and build rule-based systems by understanding input-operation-output relationships (Baron-Cohen et al. 2003).

Based on CFA results, Ling et al. (2009) found that a single-factor model for *SQ-40* has poor fit and recommended a 4-factor, 18-item solution, including *technicity, topography, DIY*, and *structure*. However, the *topography* and *DIY* subscales each had less than 3 items with "good" loadings, and overall, 8 of the 18 items have loadings that are less than "good" ($\lambda < 0.55$, Tabachnick and Fidell 2007), indicating limited convergent validity in these two subscales and the possibility that a two-factor *SQ* model (including *technicity* and *structure*) is a better fit.

Based on Manning et al.'s (2010) set of the most gender-differentiating *SQ* items, Veale and Williams (2015) tested a single-factor, 8-item measure (*SQ-8*). Though its overall model fit was thought to be "reasonably adequate" (p. 4), only 2 of its 8 items have loadings in the range of "good" or higher ($\lambda > 0.55$, Tabachnick and Fidell 2007). Therefore, further development of the *SQ* scale will likely require the higher loading items from the *SQ-8* to be supplemented by additional items from the *SQ-18* (Ling et al. 2009) to enhance scale validity.

In sum, though the existing *AQ, RBQ-2A* and *SQ* measures have been used in many empirical studies, yet they still exhibit significant psychometric issues—many of these have been acknowledged in the literature—and require additional refinement and testing to be used in future research. As the goal of this work is to rigorously evaluate, develop, and refine these frequently used scales to improve their psychometric validity and parsimony as well as ensure applicability to the general adult population, we begin with a baseline examination in Study 1 by testing the briefest existing versions of these measures before refining them in Study 2 and finally validating them in Study 3. A road map summarizing the scales, goals, and analytic techniques in each study is presented in Table 2.

Table 3 Study 1 EFA loading matrix for *AQ-10*

Factors	Item number	Item	Factor 1	Factor 2	Factor 3
Attention to detail	#5	I often notice small sounds when others do not	−0.277	0.019	0.255
	#28R	I usually concentrate more on the whole picture, rather than the small details	0.038	0.108	−0.179
Attention switching	#32R	I find it easy to do more than one thing at once	0.178	0.831	−0.089
	#37R	If there is an interruption, I can switch back to what I was doing very quickly	0.019	0.591	0.041
Communication	#27R	I find it easy to 'read between the lines' when someone is talking to me	0.612	0.339	0.024
	#31R	I know how to tell if someone listening to me is getting bored	0.585	0.138	−0.077
Imagination	#20	When I'm reading a story I find it difficult to work out the characters' intentions	0.033	0.102	0.679
	#41	I like to collect information about categories of things (e.g., types of car, train, bird, plant)	−0.078	−0.147	0.169
Social skill	#36R	I find it easy to work out what someone is thinking or feeling just by looking at their face	0.756	−0.016	0.063
	#45	I find it difficult to work out people's intentions	0.459	0.063	0.590

Notes: $N_1 = 207$; Item numbers are based off of their original number in the larger *AQ-50* instrument; R = Reverse coded item

Study 1: Initial Evaluation

Method

Participants and Procedures

An anonymous online survey was administered with a sample of undergraduate students from two large public universities in the United States. Students received extra course credit for their participation. A total of 207 students returned useable responses, including 128 males (61.8%), 73 females (35.3%), and six students who did not report their gender. The mean age was 21.86 years, with the vast majority of them (91.8%) between 20 and 25. Most respondents were majored in business (31.2%) and IT-related (62.9%) fields. A total of 11 respondents (5.3%) were international students.

Measures

In addition to the customary demographic questions (e.g., age, gender, education), the survey included the *AQ-10* (Allison et al. 2012), the *RBQ-2A* (Barrett et al. 2015), and the *SQ-8* (Veale and Williams 2015). The complete set of items is presented in Table 3 (*AQ-10*) and 4 (*RBQ-2A* and *SQ-8)*. All items were measured on a 7-point scale from "Strongly Disagree" to "Strongly Agree".

Statistical Analysis

As discussed earlier, the purpose of Study 1 is to empirically confirm the psychometric issues described in the prior literature before these measures are further developed and refined in subsequent studies. To begin our evaluation of scale validity, we first examine their internal consistency using Cronbach's α estimates, and then assess their convergent and discriminant validity with CFA using LISREL 8.80.

Results

Internal Consistency

AQ-10 As expected, all five *AQ-10* subscales exhibited low internal consistency ($\alpha_{AttentiontoDetail} = 0.21$, $\alpha_{AttentionSwitching} = 0.57$, $\alpha_{Communication} = 0.50$, $\alpha_{Imagination} = 0.21$, and $\alpha_{Social} = 0.43$). When all ten items were evaluated in a single factor, the internal consistency remained low ($\alpha = 0.48$).

RBQ-2A Both subscales, *repetitive motor behavior* and *insistence on sameness*, achieved adequate internal consistency ($\alpha_{RMB} = 0.80$ and $\alpha_{IoS} = 0.83$), and therefore were further examined in a CFA in the following section.

SQ-8 The *SQ-8* achieved acceptable internal consistency ($\alpha = 0.75$) with the existing items and was also included in the CFA test in the following section.

To sum up, in contrast to the *AQ-10*, the *RBQ-2A* and *SQ-8* were found to demonstrate adequate internal consistency (i.e., $\alpha \geq 0.70$) and were therefore further tested for validity both within and between the scales through a combined CFA.

Convergent and Discriminant Validity

Since all five *AQ-10* subscales exhibited unsatisfactory internal consistency, it was not necessary to further assess its

factorial validity using CFA in its current form. However, in case there exists an underlying model consisting of fewer factors, we explored its factor structure using an unrestricted EFA with varimax rotation (Table 3), which revealed three factors with eigen values greater than (1) However, after removing items with low loadings (#5, #28R and #41) and cross loadings (#45), only two factors remained. Factor 1 consisted of one item from *social skill* (#36) and two items from *communication*, while both items from *attention switching* loaded onto Factor (2) However, neither factor met the threshold for scale reliability (i.e., $\alpha \geq 0.70$), thus no further analysis of the *AQ-10* was necessary at this stage.

When testing the *RBQ-2A* and *SQ-8* using CFA, we first examined the two scales separately. CFA results show that the two-factor *RBQ-2A* exhibits less than satisfactory model fit ($\chi^2 = 162.98$, $df = 76$, $NFI = 0.90$, $CFI = 0.94$, $RMSEA = 0.074$ with 90% *CI*: 0.058–0.090), with both *NFI* and *CFI* below the recommended thresholds of 0.95 and *RMSEA* exceeding 0.06 for acceptable model fit (Hu and Bentler 1999).

CFA results of the single-factor *SQ-8* also indicates poor model fit ($\chi^2 = 71.27$, $df = 20$, $NFI = 0.87$, $CFI = 0.90$, $RMSEA = 0.11$ with 90% *CI*: 0.08–0.14), with *NFI, CFI* and *RMSEA* all comparing unfavorably with recommended thresholds (Hu and Bentler 1999). Additionally, only a single item loading exceeded 0.60.

After testing each scale individually, we examined the two scales together in a single model to evaluate their convergent and discriminant validity. The combined model also provided less than satisfactory fit ($\chi^2 = 373.19$, $df = 206$, $NFI = 0.84$, $CFI = 0.92$, $RMSEA = 0.063$ with 90% *CI*: 0.053 ~ 0.073). Factor loadings in Table 4 indicate that four *RBQ-2A* items (RMB #1, IoS #1, #2 and #7) have "poor" loadings ($\lambda < 0.45$, Tabachnick and Fidell 2007), which are similar to those reported by Barrett et al. (2015). Thus, these four items should be removed or refined in future steps of scale development.

The *SQ-8* item loading matrix (Table 4) suggests that Items 2, 3R, and 4R have similarly low loadings in our study as in prior examinations (e.g., Veale and Williams 2015), necessitating their removal or refinement in future steps. As also observed by Veale and Williams, Item 1 ("maps") has significant conceptual overlap with Item 6 ("motorways"), and thus could be removed to increase parsimony. Since only one item has a loading exceeding 0.60, further development of the *SQ* scale requires additional items from the larger *SQ-18* to strengthen its psychometric properties.

Discussion

Results from Study 1 indicate that, as expected, the *AQ-10* has low internal consistency in all of its five subscales compared to the established guideline ($\alpha \geq 0.70$, Nunnally 1978). Though five of its items loaded onto two factors in an unrestricted EFA, appearing to echo the current dual-factor ASC diagnostic criteria, this alternative structure for *AQ-10* cannot be supported in view of its low internal consistency, which likely resulted from its small number of items. Further development of the *AQ* requires additional items from the larger *AQ-26* (Austin 2005; Hurst et al. 2007) to supplement the *AQ-10* items before conducting further tests.

Results also show that, as expected, the *SQ-8* has low item loadings and poor overall model fit. Thus, neither the *AQ-10* nor the *SQ-8* possesses satisfactory psychometric properties for empirical survey-based research at this time. Further development and refinement of these two scales require incorporating additional items to enhance their psychometric properties.

Though the *RBQ-2A* scale also exhibits less than satisfactory model fit, its psychometric properties may be improved by simply removing items with low loadings as the number of items in each factor (6 and 8) still far exceeds the recommended minimum of three items per scale (Kline 2010). Such item removal can also further improve its parsimony. In Study 2, we refine these three scales based on the issues identified in Study 1 and use theory and prior literature to determine potential remedies for the less than satisfactory psychometric properties.

Study 2: Further Refinement

Based on results from Study 1, our focus in Study 2 was to improve scale validity by removing unsatisfactory items from the *RBQ-2A* and incorporating additional items into the *AQ* and *SQ* scales.

Method

To achieve a high degree of scale applicability and generalizability across adult populations, we collected data from a larger, more heterogeneous sample through Amazon's Mechanical Turk (MTurk). Other ASC studies have also used samples from different populations in scale development (e.g., Barrett et al. 2015; Berger et al. 2016; Odom et al. 2018). Such multi-sample design can enhance scale applicability and generalizability to a variety of settings (Hui et al. 2004), which is essential for empirical research.

The use of MTurk samples has seen significant growth in recent years in psychology, psychiatry, and other behavioral fields as a way to recruit participants that are more diverse than college students (e.g., Berger et al. 2016; Chua 2013; Gosling and Mason 2015; Longo et al. 2018; Steelman et al. 2014). MTurk samples have been found to provide highly replicated results to those of traditional

Table 4 Study 1 CFA loading matrix for *RQB-2A* and *SQ-8*

Scale	Item number	Item	RMB	IoS	SQ
Repetitive motor behavior (*RBQ-2A*, Barrett et al. 2015)	RMB_1	Do you like to arrange items in rows or patterns?	0.46	0.22	0.12
	RMB_2	Do you repetitively fiddle with items?	0.72	0.34	0.19
	RMB_3	Do you spin yourself around and around?	0.56	0.26	0.15
	RMB_4	Do you rock backwards and forwards, or side to side, either when sitting or when standing?	0.71	0.33	0.19
	RMB_5	Do you pace or move around repetitively?	0.59	0.28	0.16
	RMB_6	Do you make repetitive hand and/or finger movements?	0.82	0.39	0.22
Insistence on Sameness (*RBQ-2A*, Barrett et al. 2015)	IoS_1	Do you have any special objects you like to carry round?	0.17	0.37	0.01
	IoS_2	Do you collect or hoard items of any sort?	0.17	0.36	0.01
	IoS_3	Do you insist on things at home remaining the same?	0.34	0.72	0.03
	IoS_4	Do you get upset about minor changes to objects?	0.32	0.68	0.03
	IoS_5	Do you insist that aspects of daily routine must remain the same?	0.32	0.69	0.03
	IoS_6	Do you insist on doing things in a certain way or redoing things until they are ''just right''?	0.26	0.55	0.02
	IoS_7	Do you play the same music, game or video, or read the same book repeatedly?	0.16	0.34	0.01
	IoS_8	Do you insist on eating the same foods, or a very small range of foods, at every meal?	0.24	0.51	0.02
Systemizing Quotient (*SQ-8*, Veale and Williams 2015)	SQ8_1R	I find it difficult to read and understand maps	0.15	0.02	0.55
	SQ8_2	I find it easy to grasp exactly how odds work in betting	0.14	0.02	0.53
	SQ8_3R	I find it difficult to learn how to program video recorders	0.09	0.01	0.32
	SQ8_4R	I do not enjoy games that involve a high degree of strategy (e.g., chess, risk, games workshop)	0.12	0.02	0.45
	SQ8_5	I can remember large amounts of information about a topic that interest me, e.g., flags of the world, airline logos	0.16	0.02	0.59
	SQ8_6	I can easily visualize how the motorways in my region link up	0.14	0.02	0.53
	SQ8_7	I am fascinated by how machines work	0.20	0.03	0.75
	SQ8_8	If I were buying a stereo, I would want to know about its precise technical features	0.15	0.02	0.55

Notes: $N1 = 207$. R = reverse worded item. Highlighted items are retained for further development and testing. Those not highlighted are dropped. SQ8_1R ("maps") was removed due to its conceptual overlap with SQ8_6 ("motorways"), which was also observed by Veale and Williams (2015) and evidenced by their high correlation. With the increasing popularity of navigation devices, traditional maps are also becoming less relevant. SQ8_2 was dropped because some participants reported that they were not familiar with betting

college student samples and specialized samples through third party organizations (Mullinix et al. 2015). MTurk respondents are also less inhibited to provide truthful answers due to increased anonymity, thus reducing social desirability bias (Shapiro et al. 2013). As further assurance to data quality, we restricted participation to MTurk members with cumulative satisfaction ratings of at least 95%. Participants were offered a small monetary incentive ($1.00) to encourage their participation.

Participants and Procedures

In Study 2, an MTurk sample of 355 participants from 44 countries returned useable data, with most of them residing in India (243, 68.5%) and Venezuela (19, 5.4%). No other country represents over 5% of the sample. Most respondents are male (260, 73.2%) and young (68.7% between 18 and 34 years of age, 28.2% between 35 and 54 years of age, 3.1% are 55 or over), and have received some college education

(9.3% some college and 89.3% with bachelor's degrees or higher).

Measures

In Study 2, additional *AQ* and *SQ* items were included to enhance scale validity. The complete list of items is provided in Table 5.

AQ The six items from the *social skills* (#36 and #45), *attention to detail* (#5 and #28), and *communication/mindreading* (#27 and #31) subscales of the *AQ-10* were supplemented by 13 items from these three subscales of the larger *AQ-26* that had loadings of 0.40 or higher in prior research (Hurst et al. 2007).

RBQ-2A Due to satisfactory loadings, the ten items from Study 1 were retained.

SQ The four remaining *SQ-8* items from Study 1 (#5, #6, #7 and #8) were supplemented by six items with loadings over 0.40 from the *technicity* and *structure* subscales of the larger *SQ-18* (Ling et al. 2009). No items from its *topography* or *DIY* subscales were considered as neither factor had at least three items with sufficiently high loadings (Ling et al. 2009). Additionally, since reverse worded items can cause respondent confusion and mistakes while failing to prevent inattentive or acquiescent answering (van Sonderen et al. 2013), three such items were revised to be positively worded to increase clarity and improve internal consistency (#11 from "rarely" to "often", #43 from "I would not" to "I would", and #51 from "I do not think" to "I think").

Statistical Analysis

Our analytical approach in Study 2 follows that of Study 1 such that we examine scale reliability after first conducting CFA tests of convergent and discriminant validity to remove items with low or cross loadings. Items with significant cross loadings should be dropped regardless of their content validity because they conceptually tap more than one latent factor and are therefore theoretically ambiguous and weak in discriminant validity (Hair et al. 1998). In sum, to ensure construct validity, items with low loadings (weak convergent validity) and cross loadings (weak discriminant validity) will be removed from further consideration (Kline 2010; Nunnally 1978).

Results

Table 5 presents the CFA factor loadings for the three measurement scales in this study. We began by dropping 8 items with loadings in the "poor" range ($\lambda < 0.45$, Tabachnick and Fidell 2007), before separately testing each measure. In each individual goodness-of-fit test, we examined the modification indices to identify and remove items that have high error covariance with other items or load onto non-intended factors (Kline 2010). This process led to the removal of another 6 items (*AQ* #11 and #38; *RMB* #3; *SQ* #11, #30, and #51) to further increase the psychometric properties of each scale.

After this initial refinement of the items, the goodness-of-fit test results in Table 6 indicate that all three resulting scales achieved satisfactory model fit in their individual CFA tests, along with acceptable levels of internal consistency (all $\alpha \geq 0.70$), except for *AQ attention to detail* ($\alpha = 0.67$), which is near the threshold. Further, when assessing the three scales together in a holistic model, the model fit remained satisfactory ($\chi^2 = 499.85$, $df = 260$, $NFI = 0.93$, $CFI = 0.97$, $RMSEA = 0.051$ with 90% CI: $0.044 \sim 0.058$), indicating adequate psychometric validity and parsimony with the refined scales and items.

Discussion

After culling low-loading items from the *RBQ-2A* scale and adding high-loading items from the larger sets of *AQ-26* and *SQ-18* items, the three revised scales were subject to further testing in Study 2 using a larger, more diverse, and global, adult sample through MTurk. Results show that these item removals and additions have enhanced the internal consistency and factorial validity of the three scales.

Having used data from a global population through MTurk, the results in Study 2 provided initial evidence of scale applicability and generalizability to more diverse populations beyond the relatively homogeneous sample of U.S. college students employed in Study 1.

Finally, after finding scale structures that achieve adequate psychometric validity and parsimony in Study 2, we conducted Study 3 as a final validation of the three scales, hereinafter referred to as the *AQ-9, RBQ-2A-R*, and *SQ-7*. As a part of this final assessment, a series of additional robustness tests (i.e., measurement invariance, nomological validity) was performed to ensure their nomological validity, applicability, and consistency with prior research.

Study 3: Final Validation

As summarized in our analysis road map (Table 2), a final validation test was conducted in Study 3 to reevaluate convergent and discriminant validity of the three measures, establish scale nomological validity with known relationships from prior research, and to perform factorial invariance tests to determine scale consistency across groups. Findings of factorial validity and invariance enable the reporting of descriptive statistics, including internal consistency and

Table 5 Study 2 CFA item loadings

Scale	Item number	Item	Loading
AQ Attention to Detail (Baron-Cohen et al. 2001)	AQ_23	I notice patterns in things all the time	0.69
	AQ_6	I usually notice car number plates or similar strings of information	0.61
	AQ_19	I am fascinated by numbers	0.54
	AQ_12	I tend to notice details that others do not	0.48
	AQ_05	I often notice small sounds when others do not	0.38
	AQ_28	I usually concentrate more on the whole picture, rather than the small details	− 0.23
AQ Social Communication (Baron-Cohen et al. 2001)	AQ_27	I find it easy to 'read between the lines' when someone is talking to me	0.27
	AQ_31	I know how to tell if someone listening to me is getting bored	0.23
	AQ_36	I find it easy to work out what someone is thinking or feeling just by looking at their face	0.38
	AQ_45	I find it difficult to work out people's intentions	0.06
	AQ_38R	I am good at social chit-chat	0.87
	AQ_11R	I find social situations easy	0.87
	AQ_17R	I enjoy social chit-chat	0.89
	AQ_26	I frequently find that I don't know how to keep a conversation going	0.32
	AQ_47R	I enjoy meeting new people	0.81
	AQ_22	I find it hard to make new friends	0.48
	AQ_15R	I find myself drawn more strongly to people than to things	0.59
	AQ_35	I am often the last to understand the point of a joke	− 0.05
	AQ_44R	I enjoy social occasions	0.87
RBQ-2A Repetitive Motor Behavior (Barrett et al. 2015)	RMB_2	Do you repetitively fiddle with items?	0.67
	RMB_3	Do you spin yourself around and around?	0.67
	RMB_4	Do you rock backwards and forwards, or side to side, either when sitting or when standing?	0.70
	RMB_5	Do you pace or move around repetitively?	0.74
	RMB_6	Do you make repetitive hand and/or finger movements?	0.79
RBQ-2A Insistence on Sameness (Barrett et al. 2015)	IoS_3	Do you insist on things at home remaining the same?	0.81
	IoS_4	Do you get upset about minor changes to objects?	0.55
	IoS_5	Do you insist that aspects of daily routine must remain the same?	0.65
	IoS_6	Do you insist on doing things in a certain way or re-doing things until they are ''just right''?	0.66
	IoS_8	Do you insist on eating the same foods, or a very small range of foods, at every meal?	0.59
SQ Structure (Baron-Cohen et al. 2003)	SQ_49	I can easily visualize how the motorways in my region link up	0.55
	SQ_13	I am fascinated by how machines work	0.64
	SQ_37	When I look at a building I am curious about the precise way it was constructed	0.61
	SQ_51	When I am in a plane I think about the aerodynamics	0.46
SQ Technicity (Baron-Cohen et al. 2003)	SQ_30	I can remember large amounts of information about a topic that interest me, e.g., flags of the world, airline logos	0.54
	SQ_33	If I were buying a stereo, I would want to know about its precise technical features	0.78
	SQ_5	If I were buying a car I would want to obtain specific information about its engine capacity	0.71
	SQ_11	I rarely read articles or web pages about new technology	0.72
	SQ_20	If I were buying a computer I would want to know exact details about its hard drive capacity and processor speed	0.76
	SQ_43	If I were buying a camera I would look carefully at the quality of the lens	0.82

Notes: $N_2 = 355$. Item numbers are based off of their original number in the larger *AQ-50, SQ-50*, and *RBQ-2A* instruments. R = reverse worded item. Highlighted items are retained for further development and testing. Those not highlighted are dropped. Of the 14 items removed, 8 had "poor" loadings ($\lambda < 0.45$, Tabachnick and Fidell 2007), and 6 (*AQ* #11, #38; *RMB* #3; *SQ* #11, #30, #51) had high error covariance with other items

Table 6 Study 2 individual scale goodness-of-fit test results

Scale	Factor	Number of items	α	CFA model fit					
				χ^2	*df*	*NFI*	*CFI*	*RMSEA*	90% CI
AQ-9	Social communication	5	0.85	58.29	26	0.97	0.98	0.059	(0.039, 0.079)
	Attention to detail	4	0.67						
RBQ-2A-R	Repetitive motor behavior	4	0.82	40.03	26	0.98	0.99	0.039	(0.008, 0.062)
	Insistence on sameness	5	0.78						
SQ-7	Technicity	4	0.86	19.03	13	0.99	0.99	0.037	(0.000, 0.069)
	Structure	3	0.73						

Notes: $N_2 = 355$

factor correlations, which can be used to further assess their nomological validity.

A high degree of factorial invariance ensures that a given measurement scale is equivalent and consistent across different populations or groups. In this study, we examine gender differences as it has been repeatedly examined in prior autism research. This step is a critically important aspect of construct validity because a lack of invariance can preclude an unambiguous interpretation of between-group differences (Cheung and Rensvold 2002).

Also an important aspect of construct validity, nomological validity needs to be assessed to confirm that the focal construct is indeed correlated with other theoretically related constructs in its nomological network established in prior research (Cronbach and Meehl 1955). In this study, we assess nomological validity by examining the intercorrelations among these three autism scales and by replicating their relationships with relevant Big Five personality traits, which have been found to account for 70% of variance in autism trait scores (Schwartzman et al. 2016). The autism-Big Five linkage has also been reported in other studies (e.g., Austin 2005; Lodi-Smith et al. 2018; Rodgers et al. 2018; Schriber et al. 2014; Wakabayashi et al. 2006).

Method

Participants and Procedures

As discussed earlier, researchers in behavioral fields have increased their use of the MTurk platform to recruit broader and more heterogeneous samples than student samples. Similarly, in order to achieve greater scale applicability and generalizability in Study 3, our data was gathered from an MTurk sample, which consisted solely of U.S. participants.

A total of 442 MTurk respondents returned useable data, including 192 males (43.7%), 247 females (56.3%), and 3 participants who did not respond to the demographic questions. Compared to the global MTurk respondents in Study 2, the U.S. participants are older (41.0% between 18 and 34 years, 41.0% between 35 and 54 years, 18.0% are 55 or over) and have received less college education (33.0% some college, 58.1% bachelor's degrees or higher).

Measures

To establish nomological validity, constructs with theoretical relationships found in prior literature should be examined to identify expected relationships (MacKenzie et al. 2011). For this study, in addition to the *AQ-9, RBQ-2A-R* and *SQ-7* items from Study 2 (Table 5), scales for both neuroticism and extraversion from the Big Five Inventory (BFI, John and Srivastava 1999) were also administered for the purpose of demonstrating scale nomological validity and further evidence of convergent and divergent validity when examining the *AQ-9, RBQ-2A-R*, and *SQ-7* with additional constructs. These two BFI scales were specifically chosen in this study because they have been shown to correlate with the *AQ* and *SQ* in prior research (e.g., Austin 2005; Schwartzman et al. 2016; Wakabayashi et al. 2006). Finding a similar pattern of results with the refined *AQ-9, RBQ-2A*, and *SQ-7* scales will provide further evidence for scale validity.

Statistical Analysis

Similar to our analytical strategies in Study 2, we first use CFA tests to evaluate scale convergent and discriminant validity. We then conduct a series of factorial invariance tests to determine the consistency of the measurement scales and factors across groups (specifically males and females in this study), which should align with known gender differences in prior autism research. After establishing adequate factorial validity and invariance to enable scale estimation and interpretation, we then report descriptive statistics, including scale reliability and factor correlations (Hair et al. 1998). Finally, we establish nomological validity by examining intercorrelations among the three autism-related scales and by replicating their relationships with their known Big Five correlates.

Table 7 Study 3 CFA factor loading matrix

Item	AQ_DET	AQ_SOC	RMB	IoS	SQ_STR	SQ_TEC
AQ_23	0.82	0.00	0.33	0.30	0.52	0.33
AQ_6	0.71	0.00	0.28	0.26	0.45	0.28
AQ_19	0.69	0.00	0.28	0.25	0.43	0.28
AQ_12	0.62	0.00	0.25	0.22	0.39	0.25
AQ_17R	0.01	0.89	0.29	0.31	− 0.06	− 0.04
AQ_47R	0.01	0.84	0.28	0.29	− 0.06	− 0.04
AQ_22	0.01	0.79	0.26	0.28	− 0.06	− 0.04
AQ_15R	0.01	0.72	0.24	0.25	− 0.05	− 0.04
AQ_44R	0.01	0.90	0.30	0.32	− 0.06	− 0.05
RMB_2	0.30	0.25	0.76	0.41	0.14	0.05
RMB_4	0.28	0.23	0.70	0.38	0.13	0.05
RMB_5	0.30	0.25	0.76	0.41	0.14	0.05
RMB_6	0.30	0.24	0.74	0.40	0.13	0.05
IoS_3	0.28	0.27	0.42	0.77	0.11	0.12
IoS_4	0.27	0.26	0.41	0.75	0.11	0.11
IoS_5	0.29	0.28	0.43	0.80	0.11	0.12
IoS_6	0.25	0.24	0.37	0.69	0.10	0.10
IoS_8	0.22	0.21	0.32	0.60	0.08	0.09
SQ_49	0.35	− 0.04	0.10	0.08	0.55	0.40
SQ_13	0.50	− 0.06	0.14	0.11	0.80	0.58
SQ_37	0.46	− 0.05	0.13	0.10	0.73	0.53
SQ_33	0.32	− 0.04	0.06	0.12	0.59	0.81
SQ_5	0.30	− 0.04	0.05	0.11	0.55	0.75
SQ_20	0.28	− 0.04	0.05	0.11	0.52	0.71
SQ_43	0.25	− 0.03	0.04	0.09	0.46	0.63

Notes: $N_3 = 442$; Focal construct item loadings are highlighted. Item numbers are based off of their original number in the larger *AQ-50, SQ-50*, and *RBQ-2A* instruments

Table 8 Study 3 individual scale goodness of fit tests

Scale	Factor	Number of Items	α	χ^2	*df*	*NFI*	*CFI*	*RMSEA*	90% *CI*
AQ-9	Social communication	5	0.92	64.78	26	0.98	0.99	0.058	(0.041, 0.076)
	Attention to detail	4	0.80						
RBQ-2A-R	Repetitive motor behavior	4	0.83	76.35	26	0.97	0.98	0.066	(0.049, 0.084)
	Insistence on sameness	5	0.84						
SQ-7	Technicity	4	0.82	20.24	13	0.99	0.99	0.036	(0.000, 0.064)
	Structure	3	0.73						

Notes: $N_3 = 442$

Results

Convergent and Discriminant Validity

The CFA loading matrix for the three autism scales is presented in Table 7. Each scale exhibits satisfactory convergent and discriminant validity with items loading primarily on their focal constructs and less so on the others in the model (Kline 2010). However, the two *SQ* subscales have some slight cross-loadings, evidencing conceptual overlap and weaker discriminant validity within the *SQ* scale, but not across the other scales. Results from three separate goodness-of-fit tests also provide satisfactory evidence of psychometric validity (Table 8).

To further examine the discriminant validity between the two *SQ* subscales, we estimated a CFA model where the correlation between the two latent factors was constrained to 1. The resulting model has a poor fit ($\chi^2 = 112.21$, $df = 14$, $NFI = 0.94$, $CFI = 0.94$, $RMSEA = 0.138$), and the two-factor *SQ* model has significantly better fit ($\Delta\chi^2 = 91.97$, $df = 1$,

Table 9 Study 3 group invariance tests (male vs. female)

Model	Invariance Level	Conceptual meaning	Test statistic	*AQ-9*	*RBQ-2A-R*	*SQ-7*
1	Configural invariance	The model provides the same number of factors and the same items associated with each factor across groups (Meredith 1993)	χ^2	86.249	97.366	36.196
			df	52	52	26
			RMSEA	0.0549	0.0632	0.0424
			CFI	0.986	0.983	0.991
2	Metric invariance	The model provides the same factor loadings across groups (Cheung 2008)	χ^2	94.482	101.782	44.435
			df	61	61	33
			RMSEA	0.0501	0.0553	0.0398
			CFI	0.986	0.985	0.991
1 vs. 2			$\Delta\chi^2$	8.233	4.416	8.239
			ΔCFI	0.000	− 0.002	0.000
3	Scalar invariance	The model provides the same factor loadings and intercepts across groups (Cheung and Rensvold 2002)	χ^2	185.144	216.556	118.282
			df	77	77	43
			RMSEA	0.0802	0.0911	0.0895
			CFI	0.957	0.949	0.947
1 vs. 3			$\Delta\chi^2$	98.895	119.19	82.086
			ΔCFI	0.029	0.034	0.044
4	Complete invariance	All parameter estimates in the model are the same across groups (Cheung 2008; Cheung and Rensvold 2002)	χ^2	214.418	224.705	200.981
			df	87	87	53
			RMSEA	0.0819	0.0851	0.113
			CFI	0.950	0.948	0.892
1 vs. 4			$\Delta\chi^2$	128.169	127.339	164.785
			ΔCFI	0.036	0.035	0.099

Notes: $N_{Male} = 192$, $N_{Female} = 247$

$p < 0.00001$) despite the cross-loadings, indicating support for the two-factor model in future research.

However, researchers who remain concerned with the discriminant validity of the two *SQ* factors may alternatively only use the *technicity* subscale, which is the factor that explains the largest amount of variance in this scale in prior research (Ling et al. 2009). Similarly, in this study, *technicity* accounts for 51.4% of the total variance while *structure* explains 14.2%.

Based on CFA tests of the three scales, we again found sufficient evidence of their convergent and discriminant validity, which allowed us to further evaluate their psychometric properties through a set of factorial invariance tests.

Factorial Invariance

Factorial invariance tests were conducted to assess the extent to which each of these three measurement scales is equivalent and consistent across different groups (e.g., males and females). We examined configural invariance, metric invariance, scalar invariance, and complete invariance (Table 9), which provide increasing levels of model strictness across groups to indicate measurement consistency (e.g., Longo et al. 2017; Marques et al. 2017). Though higher levels of invariance are often hard to achieve "as metric equivalence and, particularly, scalar equivalence are frequently rejected in social science research" (Cheung and Rensvold 2002, p. 601), metric invariance is considered a prerequisite for meaningful cross-group comparison (Bollen 1989).

In this research, the factorial invariance tests were conducted between male and female participants. The choice of a gender-based assessment enables us to link our findings with existing knowledge on ASC, such as higher autistic tendencies in men than in women (e.g., Austin 2005). However, such known gender difference was also expected to make higher levels of invariance (e.g., scalar, complete) unlikely. Thus, our goal was to provide evidence of configural and metric invariance between males and females with a pattern of gender differences that are in keeping with prior literature.

In evaluating invariance, a threshold for ΔCFI of no more than 0.01 was adopted (Cheung and Rensvold 2002; Kline 2010). As shown in Table 9, all three scales showed satisfactory model fit in Models 1 and 2 and met the ΔCFI threshold, thus demonstrating evidence for configural and metric invariance. When testing for scalar and complete invariance in Models 3 and 4, model fit became less than satisfactory ($RMSEA < 0.06$, Hu and Bentler 1999) and exceeded the ΔCFI threshold as expected. This set of tests suggests that the three scales have all demonstrated configural and metric invariance, but as anticipated, not scalar or complete

Table 10 Study 3 Gender Group Comparison

Scale	Factor	Male ($N=192$)		Female ($N=247$)		t	p
		Mean	SD	Mean	SD		
AQ-9	Social communication	3.84	1.46	3.70	1.51	0.97	0.167
	Attention to detail	4.48	1.28	4.09	1.35	3.07	**0.001**
RBQ-2A-R	Repetitive motor behavior	3.00	1.32	2.83	1.45	1.22	0.111
	Insistence on sameness	3.35	1.34	3.12	1.32	1.75	**0.041**
SQ-7	Structure	4.73	1.26	3.68	1.39	8.18	**0.000**
	Technicity	5.42	1.18	4.61	1.41	6.43	**0.000**

invariance due to known gender differences also reported in prior autism research.

With evidence of metric invariance, which is a prerequisite for meaningful cross-group comparison (Bollen 1989), a gender comparison test was performed for each of the scales. As shown in Table 10, males are significantly higher than females in *attention to detail, insistence on sameness, structure*, and *technicity*. However, gender differences in *social communication* and *repetitive motor behavior* are not significant. These findings are generally expected and consistent with our invariance test results (i.e., metric, but not scalar invariance) as well as prior research showing gender differences in autistic tendencies (e.g., Austin 2005).

Descriptive Statistics and Nomological Validity

After providing evidence of adequate psychometric validity of the scales, we are able to estimate and provide interpretations of each final scale. Table 11 presents the descriptive statistics, scale reliabilities, and factor correlations for the final scales. (See complete list of the final items in the Appendix.) All scales exhibit satisfactory reliability ($\alpha > 0.70$).

As expected, many factors within the three scales are significantly correlated with one another (Table 11). As further evidence for their nomological validity, many of these factors are significantly linked to their known Big Five correlates (e.g., Austin 2005; Schwartzman et al. 2016; Wakabayashi et al. 2006). For example, *AQ Social Communication* is negatively related to extroversion ($r = -0.86$), and *RBQ-2A Repetitive Motor Behavior* is positively related to neuroticism ($r = 0.51$).

Discussion

The three measures that were further developed and refined in Study 2 were subject to a final round of testing in Study 3 to cross-validate the factor structures, psychometric properties, and nomological validity in a separate data collection (MacKenzie et al. 2011). Based on the results from a sample of U.S. MTurk members, all three scales demonstrated satisfactory internal consistency as well as convergent and discriminant validity, indicating their appropriateness for empirical survey-based research.

Further, in keeping with prior literature (e.g., Austin 2005; Schwartzman et al. 2016; Wakabayashi et al. 2006), the three scales had significant intercorrelations and replicated their relationships with the Big Five traits of extroversion and neuroticism, thus demonstrating their consistency and nomological validity (MacKenzie et al. 2011). Additionally, the factorial invariance tests indicated that the three scales exhibit satisfactory configural and metric invariance (Cheung and Rensvold 2002), which enabled us to compare means between genders, a common comparison in autism research (e.g., Austin 2005).

General Discussion

In this research, three rounds of scale testing and refinement were carried out using heterogeneous samples from general adult populations, which provided evidence of their applicability and generalizability to different populations (Kukull and Ganguli 2012). The resulting *AQ-9*, consisting of two factors: *social communication* and *attention to detail*, now mirrors the current dual diagnostic criteria in the *DSM-5* and thus better aligns academic research and clinical practice. Also containing 9 items, the *RBQ-2A-R* has been refined through CFA for the first time, providing evidence of its reliability and validity for empirical research across the general population. The *SQ-7* scale consists of two factors, including *technicity* and *structure*, and its content has also been updated to remain applicable.

To establish their psychometric properties, this research has provided evidence of scale factorial validity (through exhibiting satisfactory convergent and discriminant validity), factorial invariance (through establishing scale configural and metric invariance), nomological validity (through demonstrating consistent relationships among the three scales and with Big Five traits as in prior research), as well as parsimony (all scales contain less than ten items). These parsimonious and psychometrically satisfactory measures can provide efficient and consistent measurement of autistic

Table 11 Study 3 descriptive statistics, scale reliability and factor correlations

Factor		Number of items	Mean	SD	α	1	2	3	4	5	6	7
1	*AQ-9*—attention to detail	4	4.27	1.33	0.80							
2	*AQ-9*—social communication	5	3.76	1.49	0.92	0.01						
3	*RBQ-2A-R*—repetitive motor behavior	4	2.91	1.40	0.83	0.40**	0.33**					
4	*RBQ-2A-R*—insistence on sameness	5	3.22	1.33	0.84	0.36**	0.35**	0.54**				
5	*SQ-7*—structure	3	4.15	1.42	0.73	0.63**	− 0.07	0.18**	0.14*			
6	*SQ-7*—technicity	4	4.98	1.37	0.82	0.40**	− 0.05	0.07	0.15*	0.73**		
7	*BFI*—extroversion	8	3.82	1.39	0.91	0.01	− 0.86**	− 0.28**	− 0.32**	0.09	0.00	
8	*BFI*—neuroticism	8	3.41	1.35	0.90	0.03	0.52**	0.51**	0.40**	− 0.20**	− 0.25**	− 0.48**

Notes: $N_3 = 442$; ** $p < 0.001$, * $p < 0.01$

tendencies across various settings in future survey-based research.

However, this research is not without limitations. First, while we feel it a strength and focus of this work to achieve higher scale applicability and generalizability by validating the instruments using three different adult samples, one concern about this design is that these heterogeneous samples represent culturally distinct populations, which may result in different sociocultural expectations for appropriate behaviors and culturally different response styles across different participants, which may in turn diminish the generalizability of these instruments.[3] Therefore, the cultural validity of these three measures should be further tested to ensure their consistency and invariance across cultures as well.

Second, due to the cross-sectional design of this research, the test–retest reliability of the scales could not be assessed. While we do find consistency in the psychometric properties of our refined scales in Studies 2 and 3, future research should examine the consistency of our measures across time periods with a single sample to provide further confidence in these instruments.

Third, while these parsimonious scales demonstrate satisfactory psychometric properties, they have less comprehensive domain coverage than their respective full-length measures. However, this is less likely an issue because as reflective indicators, individual items within a scale are highly correlated and reflect a common underlying latent construct (Hair et al. 1998; Nunnally 1978). Thus, to the extent that scale reliability remains satisfactory, removing individual items should not significantly affect measurement as each item reflects the same underlying construct. Additionally, our final instruments (e.g., *AQ-9, SQ-7*) are similar in lengths to existing abbreviated measures (e.g., *AQ-10*, Allison et al. 2012, *SQ-8*; Veale and Williams 2015), which were also shortened versions of the original, larger scales (e.g., *AQ-50, SQ-40*).

Finally, though our nonclinical samples may have included some autistic individuals (either diagnosed or not), this research could have benefited from the inclusion of a separate autistic sample. The use of both neurotypical and autistic samples in future research would allow a test of the discriminant ability of these scales and potentially demonstrate their effectiveness as screening tools. While these self-report measures are useful in empirical survey-based research, they are not intended or designed to be diagnostic tools.

[3] We thank an anonymous reviewer for suggesting this point.

Conclusions

Measurement scales that are ideal for empirical survey-based research should be parsimonious and demonstrate satisfactory psychometric properties such as scale reliability and factorial validity (Kline 2010). Despite repeated usage of the *AQ, SQ*, and *RMB-2A* in the literature, there has been limited in-depth evaluation of their psychometric properties and appropriateness for empirical research. Our examination found that they could all benefit from further development.

Three rounds of scale development and refinement resulted in three psychometrically satisfactory and parsimonious instruments that can provide efficient and consistent measurement of autistic tendencies in the general adult population. These scales can increase confidence and comparability of the findings of future research while also reducing response burden.

Given that research on autism-related traits in the general adult population is an under-studied area and that the existing measures need further refinement, the development of a rigorously validated set of instruments is a meaningful contribution to this area of empirical research and clinical application.

Author Contributions All authors contributed to the design, analysis, and writing of the paper.

Compliance with Ethical Standards

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the Institutional Review Board and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent Informed consent was obtained from all individual participants included in the research.

Conflict of interest None.

Appendix: Final Scales

All items are measured on Likert-like scales of 1–7 from "Strongly Disagree" to "Strongly Agree."

All item numbers are from the original measurement instruments.

AQ-9 (Adapted from Baron-Cohen et al. 2001).

Attention to detail

6. I usually notice car number plates or similar strings of information
12. I tend to notice details that others do not
19. I am fascinated by numbers
23. I notice patterns in things all the time

Social Communication.

15R. I find myself drawn more strongly to people than to things.
17R. I enjoy social chit-chat.
22. I find it hard to make new friends.
44R. I enjoy social occasions.
47R. I enjoy meeting new people.

RBQ-2A-R (Adapted from Barrett et al. 2015).

Repetitive motor behavior

2. Do you repetitively fiddle with items?
4. Do you rock backwards and forwards, or side to side, either when sitting or when standing?
5. Do you pace or move around repetitively?
6. Do you make repetitive hand and/or finger movements?

Insistence on sameness

3. Do you insist on things at home remaining the same?
4. Do you get upset about minor changes to objects?
5. Do you insist that aspects of daily routine must remain the same?
6. Do you insist on doing things in a certain way or re-doing things until they are ''just right''?
8. Do you insist on eating the same foods, or a very small range of foods, at every meal?

SQ-7 (Adapted from Baron-Cohen et al. 2003).

Technicity

5. If I were buying a car I would want to obtain specific information about its engine capacity.
20. If I were buying a computer I would want to know exact details about its hard drive capacity and processor speed.
33. If I were buying a stereo, I would want to know about its precise technical features.
43. If I were buying a camera I would look carefully at the quality of the lens.

Structure

13. I am fascinated by how machines work.
37. When I look at a building I am curious about

the precise way it was constructed.

49. I can easily visualize how the motorways in my region link up.

References

Allison, C., Auyeung, B., & Baron-Cohen, S. (2012). Toward brief "Red Flags" for autism screening: The short autism spectrum quotient and the short quantitative checklist for autism in toddlers in 1,000 cases and 3,000 controls. *Journal of the American Academy of Child and Adolescent Psychiatry, 51*(2), 202–212. https://doi.org/10.1016/j.jaac.2011.11.003.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.). Arlington: American Psychiatric Publishing.

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. https://doi.org/10.1016/j.paid.2004.04.022.

Baron-Cohen, S. (2002). The extreme male brain theory of autism. *Trends in Cognitive Sciences, 6*(6), 248–254.

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

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*(1), 5–17.

Barrett, S. L., Uljarević, M., Baker, E. K., Richdale, A. L., Jones, C. R. G., & Leekam, S. R. (2015). The adult repetitive behaviours questionnaire-2 (RBQ-2A): A self-report measure of restricted and repetitive behaviours. *Journal of Autism and Developmental Disorders, 45*(11), 3680–3692. https://doi.org/10.1007/s10803-015-2514-6.

Berger, N. I., Manston, L., & Ingersoll, B. (2016). Establishing a scale for assessing the social validity of skill building interventions for young children with autism spectrum disorder. *Journal of Autism and Developmental Disorders, 46*(10), 3258–3269.

Bishop, S. L., & Seltzer, M. M. (2012). Self-reported autism symptoms in adults with autism spectrum disorders. *Journal of Autism and Developmental Disorder, 42*, 2354–2363. https://doi.org/10.1007/s10803-012-1483-2.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Carpenter, P. (2012). "Diagnosis and assessment in autism spectrum disorders". *Advances in Mental Health and Intellectual Disabilities, 6*(3), 121–129. https://doi.org/10.1108/20441281211227184.

Cheung, G. W. (2008). Testing equivalence in the structure, means, and variances of higher-order constructs with structural equation modeling. *Organizational Research Methods, 11*(3), 593–613.

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_5.

Chua, R. Y. J. (2013). The costs of ambient cultural disharmony: Indirect intercultural conflicts in social environment undermine creativity. *Academy of Management Journal, 56*(6), 1545–1577.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957.

Cuccaro, M. L., Shao, Y., Grubber, J., Slifer, M., Wolpert, C. M., Donnelly, S. L., et al. (2003). Factor analysis of restricted and repetitive behaviors in autism using the autism diagnostic interview-R. *Child Psychiatry and Human Development, 34*(1), 3–17. https://doi.org/10.1023/A:1025321707947.

Deutskens, E., de Ruyter, K., Wetzels, M., & Oosterveld, P. (2004). Response rate and response quality of internet-based surveys: An experimental study. *Marketing Letters, 15*(1), 21–36. https://doi.org/10.1023/B:MARK.0000021968.86465.00.

Dillman, D. A. (2000). *Mail and internet surveys: The tailored design method*. New York: Wiley.

Frazier, T. W., & Youngstrom, E. A. (2007). Historical increase in the number of factors measured by commercial tests of cognitive ability: Are we overfactoring? *Intelligence, 35*, 169–182.

Gosling, S. D., & Mason, W. (2015). Internet research in psychology. *Annual Review of Psychology, 66*, 877–902.

Guadagnoli, E., & Velicer, W. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin, 103*, 265–275.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (1998). Multivariate data analysis (Vol. 5, 3, pp. 207–219). Upper Saddle River: Prentice Hall.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118.

Hui, C., Lee, C., & Rousseau, D. M. (2004). Psychological contract and organizational citizenship behavior in China: Investigating generalizability and instrumentality. *Journal of Applied Psychology, 89*(2), 311.

Hurst, R., Mitchell, J., Kimbrel, N., Kwapil, T., & Nelson-Gray, R. (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.

Jackson, S. L. J., Hart, L., Brown, J. T., & Volkmar, F. R. (2018). Brief Report: Self-Reported Academic, Social, and Mental Health Experiences of Post-Secondary Students with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders, 48*(3), 643–650. https://doi.org/10.1007/s10803-017-3315-x.

John, O. P., & Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd edn., pp. 102–138). New York: Guilford.

Kline, R. B. (2010). *Principles and practice of structural equation modeling*. New York: The Guildford Press.

Kloosterman, P., Keefer, K., Kelley, E., Summerfeldt, L., & Parker, J. (2011). Evaluation of the factor structure of the autism-spectrum quotient. *Personality and Individual Differences, 50*(2), 310–314.

Kukull, W. A., & Ganguli, M. (2012). Generalizability: The trees, the forest, and the low-hanging fruit. *Neurology, 78*(23), 1886–1891.

Lewis, M. H., & Bodfish, J. W. (1998). Repetitive behavior disorders in autism. *Mental Retardation Research Reviews, 4*, 80–89.

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

Lodi-Smith, J., Rodgers, J. D., Cunningham, S. A., Lopata, C., & Thomeer, M. L. (2018). Meta-analysis of Big Five personality traits in autism spectrum disorder. *Autism*, https://doi.org/10.1177/1362361318766571.

Longo, Y., Coyne, I., & Joseph, S. (2018). Development of the short version of the scales of general well-being: The 14-item SGWB. *Personality and Individual Differences, 124*, 31–34.

MacCallum, R., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*(3), 490–504.

MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. *MIS Quarterly, 35*(2), 293–334.

Manning, J. T., Baron-Cohen, S., Wheelwright, S., & Fink, B. (2010). Is digit ratio (2D:4D) related to systemizing and empathizing? Evidence from direct finger measurements reported in the BBC internet survey. *Personality and Individual Differences, 48*, 767–771. https://doi.org/10.1016/j.paid.2010.01.030.

Marques, M. D., Elphinstone, B., Critchley, C. R., & Eigenberger, M. E. (2017). A brief scale for measuring anti-intellectualism. *Personality and Individual Differences, 114*, 167–174. https://doi.org/10.1016/j.paid.2017.04.001.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525–543.

Milton, D. E. M. (2012). On the ontological status of autism: the 'double empathy problem'. *Disability & Society, 27*(6), 883–887. https://doi.org/10.1080/09687599.2012.710008.

Mullinix, K. J., Leeper, T. J., Druckman, J. N., & Freese, J. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science, 2*(2), 109–138.

Nunnally, J. C. (1978). Assessment of reliability. In J. C. Nunnally (Ed.), *Psychometric theory* (2nd edn., pp. 245–246). New York: McGraw-Hill.

Nylander, L., & Gillberg, C. (2001). Screening for autism spectrum disorders in adult psychiatric out-patients: A preliminary report. *Acta Psychiatrica Scandinavica, 103*, 428–434.

Odom, S. L., Cox, A., Sideris, J., Hume, K. A., Hedges, S., Kucharczyk, S. et al. (2018). Assessing quality of program environments for children and youth with autism: Autism Program Environment Rating Scale (APERS). *Journal of Autism And Developmental Disorders, 48*(3), 913–924.

Rodgers, J. D., Lodi-Smith, J., Hill, P. L., Spain, S. M., Lopata, C., & Thomeer, M. L. (2018). Brief report: Personality mediates the relationship between autism quotient and well-being: A conceptual replication using self-report. *Journal of Autism & Developmental Disorders, 48*, 307–315. https://doi.org/10.1007/s10803-017-3290-2.

Schriber, R. A., Robins, R. W., & Solomon, M. (2014). Personality and self-insight in individuals with autism spectrum disorder. *Journal of Personality and Social Psychology, 106*(1), 112–130.

Schwartzman, B. C., Wood, J. J., & Kapp, S. K. (2016). *Journal of Autism & Developmental Disorders, 46*, 253–272. https://doi.org/10.1007/s10803-015-2571-x.

Shapiro, D. N., Chandler, J., & Mueller, P. A. (2013). Using mechanical turk to study clinical populations. *Clinical Psychological Science, 1*(2), 213–220.

Steelman, Z. R., Hammer, B. I., & Limayem, M. (2014). Data collection in the digital age: Innovative alternatives to student samples. *MIS Quarterly, 38*(2), 355–378.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th edn.). Boston: Allyn and Bacon.

van Sonderen, E., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of reverse wording of questionnaire items: Let's learn from cows in the rain. PLoS ONE 8(9): https://doi.org/10.1371/annotation/af78b324-7b44-4f89-b932-e851fe04a8e5.

Veale, J. F., & Williams, M. N. (2015). The psychometric properties of a brief version of the systemizing quotient. *European Journal of Psychological Assessment, 33*(3), 173–180. https://doi.org/10.1027/1015-5759/a000283.

Velicer, W. F., & Fava, J. L. (1998). Effects of variable and subject sampling on factor pattern recovery. *Psychological Methods, 3*, 231–251.

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. https://doi.org/10.1016/j.paid.2006.04.003.