2020

Phenotypic And Genotypic Heterogeneity In Autism Spectrum Disorder

Caitlin Crosley Clements
University of Pennsylvania

Phenotypic And Genotypic Heterogeneity In Autism Spectrum Disorder

Abstract

Many genetic events can cause autism spectrum disorder (ASD). One specific genetic event involves deletion or duplication of approximately 50 genes, 22q11.2 Deletion/Duplication Syndrome, and leads to ASD in 10-40% of cases. Chapter 1 describes an effort to identify a critical region that confers ASD risk within those ~50 genes and reports that the Low Copy Repeat-A to B region shows the strongest association. Next, we explore 'background genetics' the remainder of the genome, almost entirely inherited from one's parents - that interact with genetic events such as 22q11.2 deletions/duplications. Quantifying a heritable phenotype in one's parents can indirectly quantify the phenotype encoded in one's 'background genetics.' Heterogeneity among individuals with 22q11.2 Deletion/Duplication Syndrome, therefore, can be partially explained by heterogeneity among their parents' phenotypes. An ideal heritable trait in which to explore this framework is one of the most studied and understood constructs in psychology: IQ. However, few studies measure parental IQ due to the prohibitive cost and inconvenience of current IQ assessments. Chapter 2 reports the optimal methods for using small sample sizes to develop and calibrate a large, computer adaptive item pool for a new IQ assessment. The method described can be used to develop an online IQ test to facilitate data collection from families and understanding of 'background genetics.' Chapter 3 tests whether 'IQ' holds the same meaning for children with autism when assessed with the Differential Ability Scales, 2nd Edition (DAS-II) compared to the normative, standardization sample and reports that while verbal and nonverbal reasoning scores do function similarly between groups, the spatial composite score does not. Taken together, these three chapters advance our understanding of IQ assessment in autism and provide one example of a genetics-first sample in which these insights can be applied. Given the importance of IQ for predicting outcomes and its heterogeneity within genetically homogenous samples, the rapidly evolving field of ASD behavioral genetics stands to benefit from an efficient, valid online IQ assessment of verbal and nonverbal reasoning, which hold the same meaning for individuals with autism and typical individuals on the commonly used DAS-II.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Psychology

First Advisor

Sara R. Jaffee

Second Advisor

Robert T. Schultz

Keywords

22q11.2 deletion syndrome, 22q11.2 duplication syndrome, Autism spectrum disorder, factor analysis, intelligence, validity

Subject Categories

Psychiatric and Mental Health | Psychology

PHENOTYPIC AND GENOTYPIC HETEROGENEITY IN AUTISM SPECTRUM DISORDER

Caitlin C. Clements

A DISSERTATION

in

Psychology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2020

Supervisor of Dissertation

Robert T. Schultz

RAC Endowed Professor of Psychology

Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania

Graduate Group Chairperson

Sara R. Jaffee, Professor of Psychology

Dissertation Committee

Sara R. Jaffee, Professor of Psychology

Daniel Swingley, Professor of Psychology

John D. Herrington, Assistant Professor of Psychiatry

Dedication page

This dissertation is dedicated to my grandmother Peg Crosley, whose unconditional positive regard has helped me always to believe, on the busiest of days and the latest of nights, that I could do it.

ACKNOWLEDGMENT

I owe an overwhelming debt of thanks to many people who contributed to these three chapters and to my doctoral training. I am thankful for my committee members: Sara Jaffee, Dan Swingley, and John Herrington. My work has greatly benefitted from your feedback at each step. Thank you for your guidance throughout my training. I appreciate your time and mentorship, and each of your unique contributions on various aspects of my work over the years.

I would like to acknowledge Bob Schultz, who is simply the best advisor I could have hoped for. I felt energized after our meetings, inspired by our discussions, motivated by his gentle and optimistic edits, and secure in voicing my thoughts, knowing that he would take time to understand them, then kindly nudge them in the most successful direction. I felt happy working at the Center for Autism Research (CAR), where Bob has cultivated an egalitarian culture and light-hearted atmosphere of supporting one another and pitching in. He encouraged doorway discussions, even and especially with him and other brilliant PIs and postdocs who always made themselves available to me. Bob's support of my goals, including as they evolved and took me away from Philadelphia, allowed me to grow and flourish. Like most graduate school journeys, mine held many ups and downs, but my advisor made sure I knew that I had his support at every single obstacle and unexpected turn. For cushioning the lows and amplifying the highs, for always putting my well-being first, and for the kind and brilliant person he is, I am deeply thankful for Bob.

I would also like to thank the funding sources that made my graduate work possible: the National Science Foundation Graduate Research Fellowship, Fulbright, the Simons Foundation, the Lurie Family Foundation, the Allerton Foundation, and particularly the McMorris family. The McMorris training program seminars, presentations, and feedback were so valuable and greatly improved my ability to communicate my research.

I am particularly grateful to Ayana King-Pointer, Tiffany D'Urso Ryan, and for many years, Katie Lowe, for keeping my studies, regulatory affairs, our office space and move, and much more running so smoothly. Thank you especially for the warm way that you treat people, even when we make mistakes. Thank you from the bottom of my heart for all that you have done for me.

I would also like to thank several other individuals at CAR: Ben, for insightful comments on several manuscripts, for thinking to connect me with Marley Watkins and Kevin Antshel, for the interesting discussions along the way, expert clinical supervision, and general mentorship and support; Judi, for a highly educational and extremely enjoyable year of clinical training at 3440 and RAC, and the positive tone you set at CAR; Ashley, Brenna, and Whitney for sage advice and valued friendship; Alisa and Jen, for all that you did to bring the 22q, IQ, and meta-analysis projects to fruition from recruitment to editing, and for your good humor and friendship; to Allison, for all of your support and friendship; and to Lisa, for your sharp wit and even sharper intellect, the former of which sustained me in several less than ideal situations, and the latter of which improved several projects, and for being a most excellent lab sister especially from afar.

I am grateful to my clinical supervisors for the excellent clinical training I received: Melissa Hunt, Jenelle Nissley-Tsiopinis, Courtney Weiner, Marty Franklin, Hilary Dingfelder, Judi Miller, Tom Power, and Kevin Antshel.

Won-Chan Lee provided expertise and guidance, without which chapter two would not have been possible. Chapter three was made possible by Marley Watkins, whose experience greatly benefited my analyses and writing.

I am so thankful to Professor Mikael Landen for his gentle guidance and incisive intellect during my Fulbright fellowship and beyond. I am grateful for your unswerving support, the chance you took on me, and your continued collaboration. My year in Stockholm would have been a (statistically) significantly worse experience without Tyra Lagerberg, whose generosity, counsel, and 'cultural translation' makes her an excellent colleague and friend.

I was lucky for very fine company during graduate school. Not many people can say that they are surrounded, both near and far, by people who are very fun *and* very smart. Thank you to Becca, Leah, Eliora, Gwen, Kelly, Gabi, Bethany, Izzy, Rivka, Anika, Josh, Andrew, Kim, Claire, Ariella, Bridget, Hannah, Allie, Libbey, Jacqueline, Brede, and Christina, among others.

I also thank my beloved brothers, Will, Teddy, and Ryan. Thank you for your supportive texts and phone calls, for taking time to come visit, and for the way you love Tyler and Emma. I could not be prouder of the men you have become.

Finally, for their cheerleading during my PhD, college, high school, and everything before and in between, and especially for the uncountable sacrifices they

made for my education, I cannot sufficiently express my gratitude to my mom and dad. I owe any success that I have experienced to the foundation you sacrificed to give me. I love you immeasurably, and I am so blessed and proud to be your daughter.

To Tyler, for everything you did, and continue to do, to make this PhD and my bigger dreams possible, and to Emma – I love you both, and none of my dreams mean anything without you.

ABSTRACT

PHENOTYPIC AND GENOTYPIC HETEROGENEITY IN AUTISM SPECTRUM DISORDER

Caitlin C. Clements

Robert T. Schultz

Many genetic events can cause autism spectrum disorder (ASD). One specific genetic event involves deletion or duplication of approximately 50 genes, 22q11.2 Deletion/Duplication Syndrome, and leads to ASD in 10-40% of cases. Chapter 1 describes an effort to identify a critical region that confers ASD risk within those ~50 genes and reports that the Low Copy Repeat-A to B region shows the strongest association. Next, we explore 'background genetics' - the remainder of the genome, almost entirely inherited from one's parents - that interact with genetic events such as 22q11.2 deletions/duplications. Quantifying a heritable phenotype in one's parents can indirectly quantify the phenotype encoded in one's 'background genetics.' Heterogeneity among individuals with 22q11.2 Deletion/Duplication Syndrome, therefore, can be partially explained by heterogeneity among their parents' phenotypes. An ideal heritable trait in which to explore this framework is one of the most studied and understood constructs in psychology: IQ. However, few studies measure parental IQ due to the prohibitive cost and inconvenience of current IQ assessments. Chapter 2 reports the optimal methods for using small sample sizes to develop and calibrate a large, computer adaptive item pool for a new IQ assessment. The method described can be used to develop an online IQ test to facilitate data collection from families and understanding of

'background genetics.' Chapter 3 tests whether 'IQ' holds the same meaning for children with autism when assessed with the Differential Ability Scales, 2nd Edition (DAS-II) compared to the normative, standardization sample and reports that while verbal and nonverbal reasoning scores do function similarly between groups, the spatial composite score does not. Taken together, these three chapters advance our understanding of IQ assessment in autism and provide one example of a genetics-first sample in which these insights can be applied. Given the importance of IQ for predicting outcomes and its heterogeneity within genetically homogenous samples, the rapidly evolving field of ASD behavioral genetics stands to benefit from an efficient, valid online IQ assessment of verbal and nonverbal reasoning, which hold the same meaning for individuals with autism and typical individuals on the commonly used DAS-II.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

GENERAL INTRODUCTION

Since the 1970s, scientists understood from studying relatives with varying degrees of biological relatedness that genetics play a significant role in psychiatric disorders. Until the 2000s, however, the genes underlying autism, schizophrenia, depression, and other disorders loomed a black box. In the past decade, the etiology of Autism Spectrum Disorder (ASD) has come into focus. We now have 91 high confidence or strong candidate autism risk genes identified through exome sequencing, genome-wide association studies, and other methods (SFARI Gene Database, 2019). We are beginning to understand the relative contributions of common variants (single nucleotide polymorphisms or SNPs, which account for approximately 50% of autism risk (Gaugler et al., 2014)), and rare variants, including both small single nucleotide variants (SNVs) and large copy number variants (CNVs). We also now have an ASD polygenic risk score that quantifies an individual's autism risk from SNPs (Grove et al., 2019). We know a rare genetic event is present in approximately 10-30% of all autism cases (Vortsman et al., 2017).

Identifying specific genes associated with ASD

Given the prevalence of rare variants in autism, some of which cause genetic syndromes, it has been thought that understanding autism in genetic syndromes could generalize to an understanding of 'idiopathic' autism, or autism with unknown etiology. Autism and related phenotypes have been well characterized in several syndromes (e.g., Fragile X, Prader-Willi (15q11-q13), 16p11 Deletion Syndrome, 22q11.2 Deletion Syndrome, *CHD8*, *DRK1A*, etc.), and all characterizations describe significant

phenotypic heterogeneity. In one instance, deep probing into the biology of the gene and its different mutations uncovered a biological mechanism (different mutations resulted in opposing effects on a neuronal sodium channgel) that explained the presence or absence of an autism diagnosis among individuals with the same mutated gene, *SCN2A* (Ben-Shalom et al., 2017). Such progress in mapping phenotypic heterogeneity to specific genotypes was possible with the *SCN2A* gene, but has not yet been possible with copy number variant syndromes, which contain dozens of genes, each which may or may not contribute to the autism phenotype. Thus, after our group identified the presence of autism in a newly discovered syndrome involving ~50 genes, 22q11.2 Duplication Syndrome, and characterized vast phenotypic heterogeneity among individuals with autism and 22q11.2DupS (Wenger et al., 2016), we quickly endeavored to zero in on genes contributing to the ASD phenotype. We successfully narrowed the genetic association with ASD down to a smaller region of approximately 25 genes. This research is described in Chapter 1 (Clements et al., 2017).

Assessing IQ heterogeneity in individuals with ASD

Heterogeneity and importance of IQ. A suspected source of phenotypic heterogeneity among individuals with genetic syndromes such as 22q11.2 Deletion or Duplication Syndrome is 'background genetics,' which is the colloquial term for the remainder of the genome outside the rare CNV or other event. Almost all of the 'background genetics' are inherited from one's parents. Thus, quantifying a heritable phenotype in one's parents can indirectly quantify the phenotype encoded in one's 'background genetics.' Heterogeneity among individuals with a syndrome, therefore, can

be partially explained by heterogeneity among their parents' phenotypes. An ideal heritable trait in which to explore this framework is one of the most studied and well-understood constructs in all of psychology: IQ. Even more importantly, IQ is strongly associated with future outcomes including employment, higher education and vocational training, independent living, and quality of peer relationships (Billstedt, Gillberg, & Gillberg, 2005; Howlin, Goode, Hutton, & Rutter, 2004; Howlin, Savage, Moss, Tempier, & Rutter, 2010). Understanding causes of the heterogeneity in IQ among individuals with autism and 22q11.2Dup/DS could facilitate prediction of outcomes such as independent living. Such understanding would also be useful for 'idiopathic' ASD, as well.

Parental IQ as a determinant of child IQ and obstacles to ascertainment. IQ is a familial trait with heritability estimates of 46%-80% across the lifespan (Polderman et al., 2015) with convergence on 50% (Plomin & Stumm, 2018). Parental IQ could substantially improve our ability to predict offspring IQ and thus future outcomes, as well as improve our understanding of pleitropic effects of genes on both IQ and ASD. However, few studies include parental IQ due to major practical barriers of prohibitive cost and inconvenience.. Most current IQ assessments require in-person administration by a masters-level clinician, use of expensive materials, and usually over an hour of time. To remove these obstacles and facilitate inexpensive, remote online, self-administered IQ assessment for whole families, we designed and piloted an online computer-adaptive IQ test developed with item response theory.

Developing an alternative IQ assessment with minimal resources. We encountered major challenges in the development of this assessment, which was originally intended for use by individuals ages 6-70 of all abilities. Briefly, these challenges included lack of literature on optimal models to calibrate item parameters with 'small' sample sizes (i.e., less than 1000 participants per age group), model convergence due to the relatively large range of ability between young children and adults completing the assessment, optimizing the number and quality of common items in a nonequivalent groups anchor test (NEAT) design, choosing the optimal method of vertical score scaling to translate scores on the common IQ scale (N(100,15)) given the 'small' sample sizes and NEAT design, optimal method for linking response sets (e.g., concurrent or separate calibration), assessing effort put forth by anonymous online child and adult research participants who completed iterations of the assessment, and addressing suboptimal correlations between our assessment and gold-standard IQ tests during a small validation study, among other challenges. Many of these challenges stemmed from a deficit in the current assessment literature on 'small' (N<1000 per group) sample sizes, as many relevant studies rely on large educational datasets (e.g., state achievement tests, national college and graduate admissions tests, etc.). Chapter 2 strives to fill this gap by investigating the feasibility of developing a computer adaptive testing (CAT) item pool without extensive corporate resources. We manipulated different design and analytic methods to test the feasibility of using 300 and 500 examinees per group, and report that while using 300 examinees per group results in a high risk of failed model convergence, 500 examinees per group in combination with particular design and analytic choices can

produce acceptably low quantities of linking error in item and ability parameters. Simply put, with a specific type of data analysis, sample sizes of 500 examinees become tenable for developing a large CAT item pool.

The meaning of IQ scores in autistic individuals

The development of an IQ assessment for use with both autistic and non-autistic individuals begged the question of whether the construct of IQ holds the same meaning across these two populations. Chapter 3 answers this question by asesssing measurement invariance of a traditional IQ test, the Differential Ability Scales, Second Edition (DAS-II), in a large sample of autistic children (n=1316) compared to the normative sample (n=2000). A previous group explored a similar question and identified a social context factor in a high-functioning autism sample (Goldstein et al., 2008). We found that the DAS-II verbal and nonverbal reasoning subtests appear to hold the same meaning for the autistic sample and the normative sample, but that the spatial subtests do not. We conclude that spatial subtest scores for autistic individuals likely reflect measurement artifacts and bias.

Conclusion

Taken together, these three chapters advance our understanding of intelligence assessment in autism and provide one example of a genetics-first sample in which these insights can be applied. Given the importance of IQ for predicting outcomes and its heterogeneity within genetically homogenous samples, the rapidly evolving field of behavioral genetics in autism stands to benefit from an efficient, valid online IQ

assessment of verbal and nonverbal reasoning, which hold the same meaning for individuals with autism and typical individuals on the commonly used DAS-II.

References

Ben-Shalom, R., Keeshen, C. M., Berrios, K. N., An, J. Y., Sanders, S. J., & Bender, K. J. (2017). Opposing effects on NaV1. 2 function underlie differences between SCN2A variants observed in individuals with autism spectrum disorder or infantile seizures. *Biological Psychiatry*, *82*(3), 224-232.

Billstedt, E., Gillberg, C., & Gillberg, C. (2005). Autism after adolescence: population-based 13-to 22-year follow-up study of 120 individuals with autism diagnosed in childhood. *Journal of Autism and Developmental Disorders*, *35*(3), 351-360.

Clements, C. C., Wenger, T. L., Zoltowski, A. R., Bertollo, J. R., Miller, J. S., de Marchena, A. B., ... & Emanuel, B. S. (2017). Critical region within 22q11. 2 linked to higher rate of autism spectrum disorder. *Molecular Autism*, *8*(1), 58.

Gaugler, T., Klei, L., Sanders, S. J., Bodea, C. A., Goldberg, A. P., Lee, A. B., ... & Ripke, S. (2014). Most genetic risk for autism resides with common variation. *Nature Genetics*, *46*(8), 881.

Goldstein, G., Allen, D. N., Minshew, N. J., Williams, D. L., Volkmar, F., Klin, A., & Schultz, R. T. (2008). The structure of intelligence in children and adults with high functioning autism. *Neuropsychology*, *22*(3), 301.

Grove, J, Ripke, S, Als, T. D., Mattheisen, M, Walters, R, K., Won, H., … Borglum, A. D. (2019). Identification of common genetic risk variants for autism spectrum disorder. *Nature Genetics, 51*(3): 431-444. https://doi.org/10.1038/s41588-019-0344-8

Howlin, P., Goode, S., Hutton, J., & Rutter, M. (2004). Adult outcome for children with autism. *Journal of Child Psychology and Psychiatry, 45*(2), 212-229.

Howlin, P., Savage, S., Moss, P., Tempier, A., & Rutter, M. (2014). Cognitive and language skills in adults with autism: a 40-year follow-up. *Journal of Child Psychology and Psychiatry, 55*(1), 49-58.

Plomin, R., & von Stumm, S. (2018). The new genetics of intelligence. *Nature Reviews Genetics, 19*(3), 148.

Polderman, T. J., Benyamin, B., De Leeuw, C. A., Sullivan, P. F., Van Bochoven, A., Visscher, P. M., & Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics, 47*(7), 702.

SFARI Gene Database. (n.d.). Retrieved May 15, 2019, from https://gene.sfari.org/database/gene-scoring/

Vorstman, J. A., Parr, J. R., Moreno-De-Luca, D., Anney, R. J., Nurnberger Jr, J. I., & Hallmayer, J. F. (2017). Autism genetics: opportunities and challenges for clinical translation. *Nature Reviews Genetics, 18*(6), 362.

Wenger, T. L., Miller, J. S., DePolo, L. M., de Marchena, A. B., Clements, C. C., Emanuel, B. S., ... & Schultz, R. T. (2016). 22q11. 2 duplication syndrome: elevated rate of autism spectrum disorder and need for medical screening. *Molecular Autism, 7*(1), 27.

CHAPTER 1: CRITICAL REGION WITHIN 22Q11.2 LINKED TO HIGHER RATE OF AUTISM SPECTRUM DISORDER

This work originally appeared in Molecular Autism (2017), 8:58.

Key Words: 22q11.2 deletion syndrome, 22q11.2 duplication syndrome, Autism spectrum disorder, *RANBP1*, Screening, Atypical, Nested, Syndromic autism, Prosopagnosia, Face processing

Abstract

Previous studies have reported no clear critical region for medical comorbidities in children with deletions or duplications of 22q11.2. The purpose of this study was to evaluate whether individuals with small nested deletions or duplications of the LCR-A to B region of 22q11.2 show an elevated rate of Autism Spectrum Disorder (ASD) compared to individuals with deletions or duplications that do not include this region. We recruited 46 patients with nested deletions (n=33) or duplications (n=13) of 22q11.2, including LCR-A to B (n_{del}=11), LCR-A to C (n_{del}=4), LCR-B to D (n_{del}=14; n_{dup}=8), LCR-C to D (n_{del}=4; n_{dup}=2), and smaller nested regions (n=3). Parent questionnaire, record review, and, for a subset, in-person evaluation were used for ASD diagnostic classification. Rates of ASD in individuals with involvement of LCR-B to LCR-D were compared with Fisher's Exact Test to LCR-A to LCR-B for deletions, and to a previously published sample of LCR-A to LCR-D for duplications. The rates of medical comorbidities and psychiatric diagnoses were determined from questionnaires and chart review. We also report group mean differences on psychiatric questionnaires. Individuals with deletions involving LCR-A to B showed a 39-44% rate of ASD compared to 0% in individuals whose deletions did not involve LCR-A to B. We observed similar rates of medical comorbidities in individuals with involvement of LCR-A to B and LCR-B to D for both duplications and deletions, consistent with prior studies. Children with nested deletions of 22q11.2 may be at greater risk for autism spectrum disorder if the region includes LCR-A to LCR-B. Replication is needed.

Background

22q11.2 deletion syndrome (22q11.2DS), also known as DiGeorge or velocardiofacial syndrome, is the most common microdeletion syndrome in humans. The 3 Mb region most frequently affected in 22q11.2DS can also be duplicated, resulting in 22q11.2 duplication syndrome (22q11.2DupS; McDonald-McGinn, Emanuel, & Zackai, 1999). Previous studies suggested a prevalence of 1 in 4000 live births for 22q11.2DS, but a recent study of consecutive pregnancies found an incidence of 1 in 992 live births – similar to that of trisomy 21 (Grati et al., 2015). The incidence of 22q11.2DupS was found to be 1 in 850 (Grati et al., 2015).

The 22q11.2 region includes low copy repeats (LCRs or segmental duplication blocks) interspersed throughout the region that frequently result in "breakpoints" for a duplication or deletion. The most commonly duplicated or deleted region spans LCR-A to LCR-D (hereafter - "classic" deletion/duplication). However, smaller nested deletions occur in 15% of affected individuals and usually stretch from only LCR-A to LCR-B, or from LCR-B to LCR-D, but can also span LCR-A to LCR-C or LCR-C to LCR-D (McDonald-McGinn et al., 1999). In other cases, deletions include the area upstream of LCR-A or extend past LCR-D (see figure 1). The diagnoses of 22q11.2DS or 22q11.2DupS can refer to patients with either a classic or nested deletion/duplication.

The phenotypes of 22q11.2DS and 22q11.2DupS overlap with one another and show significant individual differences (Kobrynski & Sullivan, 2007; Wentzel, Fernström, Öhrner, Annerén, & Thuresson, 2008). The syndromes can affect almost any organ system, and individuals can present with diverse constellations of medical issues

and structural malformations, as well as a wide range of severity. Common medical comorbidities include congenital heart disease, hypocalcemia, renal abnormalities, immune deficiencies, and neuropsychiatric differences (McDonald-McGinn, Emanuel, & Zackai, 1999). There is a recognizable facial gestalt in 22q11.2DS but no recognizable gestalt has been identified in 22q11.2DupS. The rate of medical problems is much lower in 22q11.2DupS (Wenger, Miller et al.., 2016).

The 22q11.2 region has also been associated with elevated rates of autism spectrum disorder (from now on referred to as "ASD"), attention deficit/hyperactivity disorder (ADHD), and most notably, schizophrenia. A recent large study of 22q11.2DS reported psychosis in 41% of adults and ADHD in 37% of children (Schneider et al., 2014), although a psychiatric registry-based study found lower rates (Hoeffding et al., 2017). Interestingly, there are no reported individuals with 22q11.2DupS with schizophrenia, and one group even suggested that it may be protective for schizophrenia (Rees et al., 2014). In contrast, an elevated risk of ASD is found in both 22q11.2DS and 22q11.2DupS. As many as 50% of individuals with 22q11.2DS and 38% with 22q11.2DupS have received community diagnoses of autism spectrum disorder; however, fewer meet strict diagnostic criteria in research settings with reported rates of 0-18% in 22q11.2DS (Angkustsiri et al., 2014; Ousley et al., 2017; Vorstman et al., 2006) and 14-25% in 22q11.2DupS (Wenger, Miller et al., 2016).

Despite significant heterogeneity in the 22q11.2 phenotype (Michaelovsky et al., 2012), little is known about critical regions that may confer risk for any specific part of the phenotype beyond schizophrenia, cleft palate, and cardiac anomalies. Prior reports

point to *TBX1*, *CRKL*, and *MAPK1* as contributors to the cardiac (Bengoa-Alonso et al., 2016; Guo et al., 2011; Guris, Fantes, David, Druker, & Imamoto, 2001; Lindsay et al., 2001) and cleft palate phenotypes (Herman et al., 2012) in 22q11.2DS. Other research linked schizophrenia risk in 22q11.2DS to hyperprolinemia associated with lowered expression of *PRODH* (proline dehydrogenase; Jacquet et al., 2005; Raux et al., 2006). Some studies reported an association between schizophrenia risk in 22q11.2DS and the lower activity Met allele of *COMT* (catechol-O-methyltransferase; Gothelf et al., 2005; Raux et al., 2006; Vorstman et al., 2009), but larger cohort studies found no evidence (Baker, Baldeweg, Sivagnanasundaram, Scambler, & Skuse, 2005; Bassett, Caluseriu, Weksberg, Young, & Chow, 2007; Murphy, Jones, & Owen, 1999; for review, see Basset & Chow, 2008). These risk genes span the 22q11.2 region, with *COMT*, *PRODH*, and *TBX1* lying between LCR-A and LCR-B, while *CRKL* lies between LCR-C and LCD-D, and *MAPK1* lies between LCR-D and LCR-E.

Recent research identified two genes as potential mediators of the ASD risk in 22q11.2DS. Radoeva et al. reported that in a sample of 87 individuals with 22q11.2DS, individuals with ASD were more likely to carry both the low-activity alleles of *COMT* and *PRODH* (leading to high levels of proline) than individuals without ASD (Radoeva et al., 2014). Neither gene individually showed a significant direct relationship with ASD, although the pattern trended in that direction. Hidding et al. further demonstrated a quantitative relationship between ASD symptom severity and the combination of *COMT*-Met genotype and high proline levels in 45 individuals with 22q11.2DS with and without ASD (Hidding, Swaab, Sonneville, Engeland, & Vorstman, 2016). Both results suggest

that the interaction between *COMT* and *PRODH*, which lie in the LCR-A to B region, may increase ASD risk in individuals with 22q11.2DS.

The purpose of the present study was to leverage a novel study design to determine whether risk for autism can be narrowed to the LCR-A to LCR-B region within 22q11.2. Owing to the rarity of these nested structural variants, this is the first study to our knowledge that attempts to collect and phenotype large enough samples to test this hypothesis. We hypothesized that individuals harboring deleted LCR-A to LCR-B would show higher rates of ASD (Wenger, Kao et al., 2016); in addition to this region harboring *COMT* and *PRODH,* it also contains *RANBPI,* a gene involved in the metabotropic glutamate receptor (mGluR) gene network that we previously hypothesized could play a role in ASD in 22q11.2DS/DupS (Wenger, Kao et al., 2016). In addition, we describe two case studies (one from our cohort and one from the literature) with much smaller, atypical duplications within the LCR-A to B region to gain hints as to the role of specific genes.

Methods

Participants

Participants with nested 22q11.2 duplications or deletions. Participants included 43 individuals with a nested duplication (n=13) or deletion (n=30) of 22q11.2 that lay entirely within LCR-A to LCR-D but was not completely inclusive of LCR-A to LCR-D (see Table 1). The only exception to this was one participant who carried a duplication of LCR-B to LCR-D and also a very small duplication between LCR-E and LCR-F. Participants were recruited from a specialty clinic at The Children's Hospital of

Philadelphia (CHOP) or were referred from a similar specialty clinic at another institution. The CHOP "22q and You" Clinic represents the largest single-site 22q11.2 clinic in the world and maintains a large catchment area across the eastern US, with patients concentrated within a few hundred mile radius of CHOP. The sample includes probands who came to clinical attention, as well as their affected siblings (n=2 with duplication and n=3 with deletion) and parents (n=2 with duplication and n=2 with deletion) whose 22q11.2DS or 22q11.2DupS was identified after the proband's diagnostic process. The duplication or deletion was confirmed using single nucleotide polymorphism (SNP) microarray or Multiplex Ligation Probe Amplification (MLPA).

Samples whose CNVs were tested by MLPA were examined using the SALSA P250 DiGeorge diagnostic probe kit (MRC-Holland, Amsterdam, the Netherlands). Commercially available software, Gene Marker from SoftGenetics (State College, PA), was used to analyze the data. Gene Marker has developed a completely integrated application for MLPA analysis with integrated functions specific for the analysis of data derived from MLPA reactions. Samples whose CNVs were identified by SNP array were analyzed using the Affymetrix SNP Array 6.0 platform following the manufacturer's instructions (Affymetrix, Santa Clara, CA, USA). Quality control values were calculated in Affymetrix Genotyping Console (Affymetrix) and any samples with Contrast QC greater than 0.4 or mean absolute pairwise difference (MAPD) greater than 0.35 were excluded from further analysis. The B allele frequency and log R ratio plots were visualized using the Affymetrix Chromosome Analysis Suite to support CNV calls.

Three additional patients who carried very small and rare atypical duplications are included in this paper in a descriptive manner (in the Case Studies section), but are not combined with the other groups in tables, figures, or statistical analyses. One patient carried a very small duplication within LCR-A to LCR-B. The other two patients (who were related to three patients in the main LCR-B to D duplication group) carried a small duplication nested between LCR-E and LCR-F.

All 43 participants were included in the medical history chart review. Nine participants were excluded from the ASD and psychiatric symptom analyses (n=34; see Figure 2) for two types of reasons: 1) ASD classification could not be determined (n=2; see below), or 2) if they presented with another medical issue likely to affect brain development (n=2 extreme prematurity and/or birth weight $<5^{th}$ centile; n=2 with CEDNIK syndrome; n=1 with 16p11.2 deletion which is independently associated with ASD; n=2 history of hypoxic brain injury; Snyder et al., 2016; Fuchs-Telem et al., 2011; D'Angelo et al., 2016; Dudova et al., 2014). Participant characteristics of the sample excluding these 9 cases are described in Table 2. Please note that some ages differ from those in the medical record review (Table 1) because a review of updated records pertinent to ASD classification, when available, was conducted three years later to allow for infants to reach the age (3 years) at which ASD symptoms would be present.

Rates of autism were analyzed separately for individuals with nested deletions and duplications. Only one individual per family (the proband) was included to avoid confounding autism rates with risk factors shared by related individuals. In one family with B-D duplication, we included an affected family member instead of the proband

because the proband harbored a 16p11.2 deletion. For deletions, 20 individuals were included after excluding 5 parents and younger siblings (2 B-D, 2 C-D, 1 A-B; see Table 3). For duplications, 5 individuals were included after excluding 4 parents and younger siblings (4 B-D; see Table 3). No individuals presented with nested duplications involving LCR-A to LCR-B or -C.

Comparison cohorts. We compiled comparison questionnaire data from four cohorts. Detailed results of medical systems chart review, neuropsychiatric questionnaires, ASD symptoms, and adaptive functioning of these four comparison groups have been published elsewhere (Wenger, Miller et al., 2016). Two cohorts were drawn from patients at the same clinic who had a confirmed classic (LCR-A to LCR-D) 22q11.2 duplication (n=29) or deletion (n=70). A non-syndromic ASD cohort (n=70) and typically developing control cohort (n=73) were drawn from other studies of neurodevelopment at the CHOP Center for Autism Research. These four cohorts were age- and sex- matched to one another but were not as well matched to either of the small nested samples described above to allow for inclusion of all eligible individuals with a nested CNV.

Informed consent was obtained for all 22q11.2 participants, as well as for all participants in the comparison cohorts (Institutional Review Board protocols #13-101307, #09-007275, #07-005689, #10-007622).

Procedures

We collected data from record review, questionnaires administered remotely, and, for a subset, an autism-specific evaluation. Record review included the participant's

electronic health record at CHOP whenever possible, as well as external medical and educational records (e.g., IEP evaluations) provided by families for individuals who did not receive routine medical care at our institution.

Medical Record and Developmental History Review. Medical and developmental history was obtained from a questionnaire completed by the participant. A licensed pediatrician and medical geneticist (TLW) reviewed clinic notes, progress reports, radiology reports, laboratory reports, etc. in each participant's record to confirm key components reported by participants. Psychiatric and neurodevelopmental diagnostic history was documented in this process as it is routinely collected during clinical visits. Families were contacted by phone to resolve questions or discrepancies.

ASD diagnostic classification.

Sources of diagnostic information. Given that our hypotheses concerned rates of ASD, particular care was given to the ASD classification process. We assigned diagnostic status after a thorough record review of clinical, research and educational records provided by families and available in the CHOP electronic health record. Participants differed in the frequency with which they received documented CHOP care. Continuous longitudinal data from CHOP developmental pediatricians and psychiatrists existed for individuals who lived locally, whereas records of individuals who lived further away or moved sometimes contained only the initial "22q and You" clinic evaluation. Participants were also asked to provide external medical and educational records.

All families were invited for an in-person ASD evaluation using the Autism Diagnostic Observation Schedule (ADOS and ADOS-2), parent interview, and IQ testing

to complete a DSM-5 (Diagnostic and Statistical Manual of Mental Illness, 5th edition) checklist (American Psychiatric Association, 2013; Lord, Rutter, DiLavore, & Risi, 1999). However, since many of our families lived far away, this proved unfeasible for a large percentage of the cases. Families who could not complete an in-person evaluation were invited for an hour-long parent phone interview with a clinician asking follow up questions to Social Communication Questionnaire, Lifetime (SCQ) responses to complete an accurate DSM-5 checklist (Rutter, Bailey, & Lord, 2003).

"ASD" group. We assigned participants to the "ASD" group if there was documentation of an ASD diagnosis (n=5 deletions, n=1 duplications). Five individuals had a diagnostic evaluation in their record; one did not, but had frequent references to the ASD diagnosis throughout the record. All participants scored above threshold (15) on the SCQ.

"No ASD" group. We assigned "No ASD" (n=20 deletions, n=8 duplications) if ASD had been considered but specifically ruled out (n=13 deletions, n= 3 duplications), or if there was no indication of ASD concerns in the available records (n=7 deletions, n=5 duplications). Two individuals (both LCR-B to D deletions) were excluded because a referral for an ASD evaluation had been recommended recently but not completed.

The *absence* of parental or professional concern about ASD is not routinely documented. Thus, we further investigated this group to determine whether there was a true absence of concern, or a lack of information. We studied parent/spouse report, provider report, behavioral descriptions, and referral history. The 22q clinic routinely refers to developmental behavioral pediatrics or psychiatry if parents indicate relevant

concerns during intake, but parents without those concerns would not have had these appointments scheduled. The 22q clinic also routinely questions parents regarding developmental history and previous concern of psychiatric diagnoses from school or medical professionals, as this population is at high risk for psychosis and other psychiatric disorders. Any concerns and prior assessments are documented in detail. Therefore, we feel confident that families were routinely asked about developmental concerns, and thus that a lack of referrals and text about concerns was a reasonably robust indicator of a lack of ASD concerns.

Neuropsychiatric questionnaires. We collected neuropsychiatric questionnaire data from participants under age 18. Questionnaires included a measure of adaptive functioning (Vineland Adaptive Behavior Scales – 2nd Edition, "Vineland-II," completed for participants 0-18 years old; Sparrow, Cicchetti, & Balla, 2005), a screener for psychiatric disorders based on DSM-IV checklists (Child and Adolescent Symptom Inventory-4R, "CASI-4R," completed for participants 5-18 years old depending on disorder; Gadow & Sprafkin, 2005), and two measures of social behavior and autistic symptoms (Social Communication Questionnaire - Lifetime, "SCQ," completed for participants 4 and above (Rutter, Bailey, & Lord, 2003), and the Social Responsiveness Scale or Social Responsiveness Scale, 2nd edition, "SRS-2" for participants 2.5-18 years old; Constantino & Gruber, 2012a; Constantino & Gruber, 2012b). Every questionnaire offers excellent psychometric properties and all but the Social Communication Questionnaire provide standardized scores based on a large, representative norming

sample. Please see Appendix, Tables s1-s4 for characteristics of subsamples that completed each questionnaire.

Analysis.

ASD rate. To test our hypothesis that the LCR-A to LCR-B region might confer increased risk of ASD in 22q11.2 duplication and deletion syndromes, we compared ASD rates among individuals whose deletion affected the LCR-A to LCR-B region ("AB/AC group:" LCR-A to B, or LCR-A to C) to individuals whose deletion did not affect the LCR-A to LCR-B region ("BD/CD group:" LCR-B to D, and LCR-C to D). Thus, our first analysis compared the "AB/AC group" to the "BD/CD group" for deletions only. In a second, more conservative analysis, we compared only individuals with deletions of LCR-A to B to those with LCR-B to D (excluding cases with deleted LCR-A to C or LCR-C to D) to match the groups on approximate size and number of genes in the deletion.

Rates were compared using a one-tailed Fisher's Exact Test to account for cells with $n<5$. An odds ratio (OR) cannot be computed when certain cells contain 0 observations; in these cases, we present 95% confidence intervals and *p* values from Fisher's Exact Test and effect sizes as chi-square statistics.

Our sample included no individuals with nested duplications involving LCR-A to B (*i.e.*, no "AB/AC" group for duplications). Thus, we compared the BD/CD duplication group to individuals with the classic LCR-A to D duplication, which does involve LCR-A to B. These results are provided for descriptive purposes only due to the sample size of

the nested duplications, which although is one of the largest reported, remains quite small.

Psychiatric symptoms (standardized questionnaires). In our dimensional analysis of psychiatric symptoms using questionnaire data, we analyzed raw scores on the SCQ, age-normed scores on the Vineland-II and SRS-2, and symptom composite scores on the CASI-4R. For deletions, we compared individuals in the "BD/CD" group to the "AB/AC" group. For duplications, we compared individual in the "BD/CD" group to the comparison cohort of classic duplications because our sample included no AB/AC duplications. We also compare the "AB/AC" deletion group to the classic deletion group as this information might prove directly useful clinically. Our interpretations focus on the size of the effect and its confidence interval, as opposed to inferential statistics, to avoid making overly strong statements based on a small sample, as suggested by many recent position papers, *e.g.*, Button et al. (2013) and Cumming et al. (2014). We present the effect sizes for each analysis and make our data available upon request so that the data generated here can be leveraged in any future meta-analyses to test our hypothesis directly.

Medical and psychiatric diagnoses. We present rates of psychiatric and medical comorbidities by nested region separately for individuals who did and did not receive recommended screening. All analyses are descriptive and for characterization purposes only. Statistical significance was not tested due to small sample sizes within each nested region.

Results

Higher rates of ASD when LCR-A to B involved

We observed a trend toward a higher rate of ASD among probands with deletions in the AB/AC group (41.7%, or 5 in 12 individuals with LCR-A to B, or LCR-A to C) compared to the BD/CD group (0%, or 0 in 8 individuals with LCR-B to D, or LCR-C to D; χ= 4.4, p=0.051, CI: 0.99, Inf; see Table 3). In a more conservative analysis that matched groups on approximate size of deleted region, we continued to observe similar rates of ASD within each group (44.4%, or 4 of 9 individuals with deletions of LCR-A to B, and 0%, or 0 in 6 individuals with deletions of LCR-B to D; χ=3.64, p=0.092, CI: 0.702, Inf). The rate of ASD did not change meaningfully when related individuals were included to increase sample size; the increased sample size provided more statistical power and revealed significant results (n=25; 38.5% rate in AB/AC group, 0% in BD/CD group; χ=5.77, p=0.024, CI: 1.39, Inf). Thus, the LCR-A to B region may confer increased risk of ASD diagnosis but a larger sample without related individuals is needed to confim.

Among duplications, individuals with the classic and BD/CD duplications showed similar rates of ASD (24.1% rate or 7 of 29 in classic group, 20% rate or 1 of 5 in BD/CD; OR=0.79, p=0.764, CI: 0.03, Inf). Results did not change meaningfully when related individuals were included to increase sample size (21.4% rate in classic group, 11.1% rate or 1 of 9 in BD/CD; OR=0.40, p=0.65, CI: 0.02, Inf), but this analysis in particular would benefit from a larger sample.

Our categorical analysis was supported by quantitative reports of autistic symptoms in the SRS-2 and SCQ (see Figure 3). A subset of each group (BD/CD deletions, AB/AC deletions, BD/CD duplications, classic duplications, classic deletions) completed the SCQ, including both individuals with and without ASD diagnoses. For deletions, the BD/CD group showed less autistic symptoms than the AB/CD group with large effect sizes (*d*'s of 1.01 and 1.20). For duplications, the difference was small-to-medium (*d*'s of 0.27 and 0.50) between the BD/CD group and the classic group. No effects reached statistical significance (see Table 5).

Moderately lower adaptive and social functioning when AB region involved

We computed effect sizes for differences in autistic symptoms, psychiatric symptoms, and adaptive behavior skills (see figures 2 and 3, table 4, additional file 2). For duplications, the differences were usually small between the "BD/CD" group and the classic duplication group (see Table 4, "Classic Duplication" rows). For deletions, the "BD/CD" group showed less impairment than the "AB/AC" group across most measures with medium or large effect sizes that did not reach statistical significance. We also calculated effect sizes for group differences between the AB/AC deletions and classic AD deletion groups and observed small or medium differences (see Table 4, "Classic Deletion" rows). We observed negligible differences between these two groups on most adaptive functioning scales. The classic deletion group showed slightly lower levels of autistic symptoms compared to the AB/AC group – small to medium effect sizes on the SRS-2 and SCQ – that were not statistically significant.

Increased rates of psychiatric disorders

In individuals with nested duplications or deletions, we observed elevated rates compared to population means in nearly every psychiatric disorder reported, including ADHD, OCD, mood dysregulation disorders, ODD and related behaviors, depression, language disorders, global developmental delay, and intellectual disability. See table 5 for observed rates of disorders by type of nested deletion or duplication.

Higher rates of medical comorbidities

We documented presence or absence of having received an appropriate screening test, and whether or not an abnormality was identified, in individuals with nested deletions and duplications between LCR-A and D (see Table 6). In order to calculate conservative estimates for the prevalence of each medical comorbidity in each group, we report both the percentage of screened individuals and the percentage of total individuals.

Case Study 1

Isolating specific genes: An individual with ASD and tiny duplication involving *RANBP1* and *COMT*, not *TBX1*. One individual in our sample came to attention of clinical geneticists due to autism spectrum disorder and was found to have a small, 300kb microduplication within the LCR-A to B region that included *RANBP1* and *COMT* but not *TBX1*. Detailed clinical evaluation and all recommended medical screening for individuals with 22q11.2 related disorders revealed none of the medical issues or dysmorphic features characteristic of the syndrome. However, the individual met diagnostic criteria for ASD, anxiety, and ADHD after evaluation by a neurodevelopmental pediatrician and standardized neuropsychiatric evaluation. The inheritance of this microduplication is unknown because paternal testing was not

possible. To our knowledge, no relatives carry an autism diagnosis but none have received formal evaluation. The individual's SNP array showed no other pathogenic variants. This individual was not included in group analyses because the duplication did not encompass the full LCR-A to B region.

Case Study 2

The role of background genetics: a family with LCR-B to D duplication and distal E-F duplication and autism and face processing deficits. The only individual in our analyses with autism in the BD/CD group carried a duplication of LCR-B to D. She had one sibling with the same LCR-B to D duplication, and two siblings with a duplication of *TOP3B* (in a small region between LCR-E and F). One of the siblings with the *TOP3B* duplication had a history of an autism diagnosis but did not currently present with significant autism symptoms. Furthermore, the proband and the sibling with LCR-B to D duplication both showed decreased face processing abilities on the Benton Facial Recognition Test (mildly impaired in the proband, clinically impaired in the sibling). Face processing difficulties have not been reported in 22q syndromes before, and we do not posit that they are central to the syndromes, but rather that the family history of possible ASD and the genetic complexity of the family raises the question that other genetic factors may have contributed to the proband's autism. Future studies of autism in nested 22q11.2 should evaluate family members for ASD, and evaluate probands for phenotypes seen in other family members, to better understand the contribution of background genetics.

Discussion

To our knowledge, this study includes the largest group of individuals with nested deletions and duplications of 22q11.2 to be compared prospectively to classic deletions and duplications with standardized measures. These data suggest that individuals with deletion of the LCR-A to B region may have a higher rate of ASD (39-44%) than those without involvement (0%); the pattern was not replicated for duplications. Taken in conjunction with Case Study 1, these findings are consistent with our hypothesis that LCR-A to B may confer risk for ASD in 22q11.2 related disorders. However, we offer this evidence as preliminary support that requires further exploration with additional samples.

It is notable that the nested deletions of all individuals with ASD involved LCR-A to B, and that we observed negligible differences between this group and the classic deletion spanning LCR-A to D in adapative functioning. These results suggest that LCR-A to B could be contributing to the autistic phenotype in individuals with classic 22q11.2DS, as well as to decreased adaptive functioning. It is also notable that we observed no duplications of LCR-A to B or LCR-A to C in our full sample of 43 individuals, although such individuals are mentioned in much larger studies (Hadley et al., 2014). Thus, it remains to be tested in larger samples whether these individuals are as likely to present with ASD as those with the classic A-D duplication.

Implications for medical screening

Prior studies have suggested that individuals with nested deletions have similar types of medical problems to those with classic deletions and should receive similar

clinical treatment. The medical chart review of our patients supported this hypothesis. It also suggested that our patients are representative of other previously reported patients with nested deletions with regard to the frequency and types of medical problems. It is notable that there appeared to be fewer medical problems in individuals with LCR-C to D. However, this region is much smaller, encompassing fewer genes than the other regions. In size and total number of genes, LCR-A to LCR-B and LCR-B to LCR-D are roughly equivalent, and the rates of medical comorbidities are similar. We also observed higher rates of some medical comorbidities in several of the nested groups as compared to individuals with full LCR-A to LCR-D deletions (*e.g.*, cervical spine anomalies in 100% of screened individuals with LCR-A to LCR-B deletion), but our sample sizes are too small to determine if this is due to chance or truly represents a higher risk subgroup. We were somewhat surprised to find that many patients had not completed portions of the recommended medical screening for individuals with 22q11.2 related disorders. It is unclear if this is due to a perception by providers that individuals with nested deletions do not need as aggressive screening as those with full deletions or duplications. Overall, we observed rates of each of the medical comorbidities in the LCR-A to LCR-B and LCR-B to LCR-D subgroups that are comparable to rates in individuals with full LCR-A to LCR-D deletions or duplications. Although the rate of medical problems appears lower in the LCR-C to LCR-D deletion and duplication groups, the sample sizes are extremely small, and therefore no strong conclusions can be made about the validity of an altered screening protocol for these patients.

***RANBP1* as a potential ASD candidate gene**

The LCR-A to B region associated with ASD risk in our sample involves approximately 25 genes, including *COMT*, *PRODH*, and *TBX1*. Prior research implicates the interaction of low activity *COMT* and *PRODH* alleles in ASD risk (Radoeva et al., 2014; Hidding et al., 2016). Other genes in the region may also confer ASD risk, and indeed the risk could be additive. We propose another possible candidate gene, Ran-binding protein 1 (*RANBP1)*, which could not be examined given our study design and might warrant further investigation. We base this speculation on five circumstantial pieces of evidence.

First, we cite the involvement of *RANBP1* in the metabotropic glutamate receptor (mGluR) gene network (Hadley et al., 2014), which is disrupted in two other syndromic forms of ASD, fragile X syndrome and tuberous sclerosis complex (Auerbach, Osterweil, & Bear 2011). Second, we previously observed a 10-fold increase in ASD rate among individuals with 22q11.2DS with a "second hit" in an mGluR network gene compared to individuals without a "second hit" (5 affected in 25 individuals with 22q11.2 compared to 1 in 50; Wenger, Kao et al., 2016). Third, two teratogens associated with increased rates of ASD – valproate and thalidomide – both decrease expression of *RANBP1* (Christinen et al., 2013; Ingram, Peckham, Tisdale, & Rodier, 2000; Meganathan et al., 2012). Fourth, the important link between *RANBP1* and expression in human brains was demonstrated by Meechan et al. (2006), who showed higher *RANBP1* expression in developing fetal brains compared to adult brains during a peak in neurogenesis. Finally, several studies in the 22q11.2 animal literature highlight *RANBP1* as important for neural development in 22q11.2 (e.g., Meechan et al., 2006; Meechan, Tucker, Maynard, &

LaMantia, 2009; Paronett, Meechan, Karpinski, LaMantia, & Maynard, 2014). Taken together, these disparate pieces of literature converge on a role of *RANBP1* in brain development, and potentially in ASD. Like other genes and gene families recently associated with ASD, *RANBP1* serves a general function within the cell (metabolizing GTP and regulating material transport to the nucleus; Zhang, Arnaoutov, & Dasso, 2014). *RANBP1* has not been identified previously as an ASD candidate gene in large ASD studies; of the approximately 25 genes in the 22q11.2 LCR-A to LCR-B region, previous genome-wide association studies or whole exome sequencing studies have idenfied *PRODH* as a candidate gene with suggestive evidence and *TBX1* and *GNB1L* as candidate genes with minimal evidence at this time (SFARI gene database https://gene.sfari.org/database/human-gene/). It is not yet clear whether genes in this region modify ASD risk in the general population, or in the context of 22q11.2 syndromes alone.

Insights from two case studies involving *TBX1* and *RANBP1*

Individuals with very small nested duplications and deletions offer a unique method of studying the associations between isolated regions or genes and individual features of the 22q11.2DS phenotype. In the present study, we could not tease apart the contributions of individuals genes to portions of the phenotype, as the LCR-A to B region includes 25 genes. Here we contrast two case studies, Case Study 1 and a prior case study by Weisfeld-Adams and colleauges (Weisfeld-Adams, Edelmann, Gadi, & Mehta, 2012), with a very small duplication including either *TBX1* or *RANBP1*, but not both, to provide some insight into the possible relative contributions of *TBX1* and *RANBP1* to the

phenotype in a descriptive fashion. Weisfeld-Adams et al. described a patient and sibling with duplication of six genes including *TBX1* but not *RANBP1*. This proband showed complex medical problems, but neither the 19-month-old proband nor the 3-year-old sibling showed any symptoms of autism or neurodevelopmental delay besides mild motor delay. (Although no concern for ASD was noted at 19 months of age, we caution against over-interpretation because ASD can be missed in toddlers when symptoms are not severe. However, by 19 months of age most children with 22q11.2DS show significant delays, little speech, and aloof social behavior, so the lack of delay suggests social development was on course.) In contrast, in Case Study 1 we described an individual with microduplication involving *RANBP1* but not *TBX1* who had ASD but no medical comorbidities. Both our patient, who had a purely psychiatric phenotype and duplication that *does* involve *RANBP1*, and the case presented by Weisfeld-Adams et al. – a purely medical phenotype that does *not* involve *RANBP1* – provide preliminary suggestive evidence that *RANBP1*, not *TBX1*, specifically might confer risk for ASD and other psychiatric diagnoses. Both microduplications include *COMT* and exclude *PRODH*, so we cannot speculate about the roles of these genes based on case studies.

Limitations

The two primary limitations of our study lie in the phenotyping and the sample size. This single-site study relied primarily on questionnaires and chart review, supplemented by in-person evaluation when feasible for the family. Thus the phenotyping, while accurate, could be improved with systematic prospective evaluations. Our sample size was small, owing to the rarity of individuals with nested duplication or

deletions in the 22q11.2 region. Our study would benefit from replication with a multi-site study that combines clinics around the world to improve statistical power.

Another limitation includes the unknown role of background genetics. We were unable to account for other contributors to ASD risk, such as common variants or known pathogenic variants occurring outside 22q11.2 that would be identified with whole exome sequencing, not clinical genetic testing with MLPA and SNP arrays. However, this risk is likely to affect all groups equally. Furthermore, we believe this unknown potential risk is likely to be small compared to the known, larger ASD risk of carrying 22q11.2DS or DupS.

Future directions might involve whole-exome sequencing of 22q11.2 samples to identify other factors that contribute to ASD risk. Such a study should include an analysis leveraging the sequencing of *PRODH*, *COMT*, *RANBP1*, and *TBX1* in individuals with nested 22q11.2 deletions and duplications to isolate the influence of these mutations on the ASD phenotype.

Conclusions

We present data on medical and psychiatric issues in 44 individuals with nested duplications and deletions within the LCR-A to D region, along with two additional siblings with tiny duplication of *TOP3B,* the largest cohort of this type to be studied prospectively. We found increased rate of ASD among individuals with deleted LCR-A to B, compared to individuals whose nested deletions did not involve that region. We tentatively speculate that *RANBP1* could provide a potential mechanistic explanation for increased rates of ASD based on this finding, our reported case study, environmental

ASD risk factors that also alter *RANBP1* expression, *RANBP1*'s role in the mGluR network, and the role of the mGluR network in other syndromic forms of ASD. We also conclude from our observation of the full spectrum of medical issues in each group that at this time, there is insufficient evidence to limit medical screening in individuals with nested duplications or deletions within the 22q11.2 region.

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders, 5th ed.* Arlington, VA: American Psychiatric Publishing.

Auerbach, B. D., Osterweil, E. K., & Bear, M. F. (2011). Mutations causing syndromic autism define an axis of synaptic pathophysiology. *Nature*, *480*(7375), 63–68. https://doi.org/10.1038/nature10658

Bassett, A. S., & Chow, E. W. C. (2008). Schizophrenia and 22q11.2 deletion syndrome. *Current Psychiatry Reports*, *10*(2), 148–157.

Bengoa-Alonso, A., Artigas-Lopez, M., Moreno-Igoa, M., Cattalli, C., Hernandez-Charro, B., & Ramos-Arroyo, M. A. (2016). Delineation of a recognizable phenotype for the recurrent LCR22-C to D/E atypical 22q11.2 deletion. *American Journal of Medical Genetics Part A*, *170*(6), 1485–1494. https://doi.org/10.1002/ajmg.a.37614

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Christensen, J., Gronborg, T. K., Sorensen, M. J., Schendel, D., Parner, E. T., Pedersen, L. H., & Vestergaard, M. (2013). Prenatal Valproate Exposure and Risk of Autism Spectrum Disorders and Childhood Autism. *JAMA*, *309*(16), 1696. https://doi.org/10.1001/jama.2013.2270

Clelland, C. L., Read, L. L., Baraldi, A. N., Bart, C. P., Pappas, C. A., Panek, L. J., … Clelland, J. D. (2011). Evidence for association of hyperprolinemia with schizophrenia and a measure of clinical outcome. *Schizophrenia Research, 131*(1–3), 139–145. https://doi.org/10.1016/j.schres.2011.05.006

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

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

Constantino, J., & Gruber, C. (2012). *The Social Responsiveness Scale -- Second Edition (SRS-2)*. Los Angeles: Western Psychological Services.

Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science, 25*(1), 7–29. https://doi.org/10.1177/0956797613504966

D'Angelo, D., Lebon, S., Chen, Q., Martin-Brevet, S., Snyder, L. G., Hippolyte, L., … for the Cardiff University Experiences of Children With Copy Number Variants (ECHO) Study, the 16p11.2 European Consortium, and the Simons Variation in Individuals Project (VIP) Consortium. (2016). Defining the Effect of the 16p11.2 Duplication on Cognition, Behavior, and Medical Comorbidities. *JAMA Psychiatry*, *73*(1), 20. https://doi.org/10.1001/jamapsychiatry.2015.2123

Developmental Disabilities Monitoring Network Surveillance Year 2010 Principal Investigators, & Centers for Disease Control and Prevention (CDC). (2014). Prevalence of autism spectrum disorder among children aged 8 years - Autism

and developmental disabilities monitoring network, 11 sites, United States, 2010. *Morbidity and Mortality Weekly Report: Surveillance Summaries, 63*(2), 1–21.

Duarte, M., Afonso, J., Moreira, A., Antunes, D., Ferreira, C., Correia, H., … Sequeira, S. (2017). Hyperprolinemia as a clue in the diagnosis of a patient with psychiatric manifestations. *Brain & Development, 39*(6), 539–541. https://doi.org/10.1016/j.braindev.2017.01.008

Dudova, I., Markova, D., Kasparova, M., Zemankova, J., Beranova, S., Urbanek, T., & Hrdlicka, M. (2014). Comparison of three screening tests for autism in preterm children with birth weights less than 1,500 grams. *Neuropsychiatric Disease and Treatment*, 2201. https://doi.org/10.2147/NDT.S72921

Elliott, C. (2007). *Differential Ability Scales - II*. London: Pearson.

Fuchs-Telem, D., Stewart, H., Rapaport, D., Nousbeck, J., Gat, A., Gini, M., … Sprecher, E. (2011). CEDNIK syndrome results from loss-of-function mutations in SNAP29: CEDNIK syndrome. *British Journal of Dermatology, 164*(3), 610-616. https://doi.org/10.1111/j.1365-2133.2010.10133.x

Gadow, K., & Sprafkin, J. (2005). *Child and adolescent symptom inventory-4R*. Stony Brook: Checkmate Plus.

Grati, F. R., Molina Gomes, D., Ferreira, J. C. P. B., Dupont, C., Alesi, V., Gouas, L., … Vialard, F. (2015). Prevalence of recurrent pathogenic microdeletions and microduplications in over 9500 pregnancies. *Prenatal Diagnosis, 35*(8), 801–809. https://doi.org/10.1002/pd.4613

Guo, T., McDonald-McGinn, D., Blonska, A., Shanske, A., Bassett, A. S., Chow, E., … International Chromosome 22q11.2 Consortium. (2011). Genotype and cardiovascular phenotype correlations with TBX1 in 1,022 velo-cardio-facial/DiGeorge/22q11.2 deletion syndrome patients. *Human Mutation*, *32*(11), 1278–1289. https://doi.org/10.1002/humu.21568

Guris, D. L., Fantes, J., Tara, D., Druker, B. J., & Imamoto, A. (2001). Mice lacking the homologue of the human 22q11.2 gene CRKL phenocopy neurocristopathies of DiGeorge syndrome. *Nature Genetics*, *27*(3), 293–298. https://doi.org/10.1038/85855

Hadley, D., Wu, Z., Kao, C., Kini, A., Mohamed-Hadley, A., Thomas, K., … Hakonarson, H. (2014). The impact of the metabotropic glutamate receptor and other gene family interaction networks on autism. *Nature Communication*, *5*, 4074.

Herman, S. B., Guo, T., McGinn, D. M. M., Blonska, A., Shanske, A. L., Bassett, A. S., … and the International Chromosome 22q11.2 Consortium. (2012). Overt cleft palate phenotype and *TBX1* genotype correlations in velo-cardio-facial/DiGeorge/22q11.2 deletion syndrome patients. *American Journal of Medical Genetics Part A*, *158A*(11), 2781–2787. https://doi.org/10.1002/ajmg.a.35512

Hoeffding, L. K., Trabjerg, B. B., Olsen, L., Mazin, W., Sparsø, T., Vangkilde, A., … Werge, T. (2017). Risk of psychiatric disorders among individuals with the 22q11.2 deletion or duplication: A Danish nationwide, register-based study.

JAMA Psychiatry, *74*(3), 282–290. https://doi.org/10.1001/jamapsychiatry.2016.3939

Ingram, J. L., Peckham, S. M., Tisdale, B., & Rodier, P. M. (2000). Prenatal exposure of rats to valproic acid reproduces the cerebellar anomalies associated with autism. *Neurotoxicology and Teratology*, *22*(3), 319–324.

Jerome, L. A., & Papaioannou, V. E. (2001). DiGeorge syndrome phenotype in mice mutant for the T-box gene, Tbx1. *Nature Genetics*, *27*(3), 286–291. https://doi.org/10.1038/85845

Lindsay, E. A., Vitelli, F., Su, H., Morishima, M., Huynh, T., Pramparo, T., … Baldini, A. (2001). Tbx1 haploinsufficieny in the DiGeorge syndrome region causes aortic arch defects in mice. *Nature*, *410*(6824), 97–101. https://doi.org/10.1038/35065105

Lord, C. (2008). *Autism Diagnostic Observation Schedule*. Los Angeles: Western Psychological Services.

Meechan, D. W., Maynard, T. M., Wu, Y., Gopalakrishna, D., Lieberman, J. A., & LaMantia, A. S. (2006). Gene dosage in the developing and adult brain in a mouse model of 22q11 deletion syndrome. *Molecular and Cellular Neuroscience*, *33*(4), 412-428.

Meechan, D. W., Tucker, E. S., Maynard, T. M., & LaMantia, A. S. (2009). Diminished dosage of 22q11 genes disrupts neurogenesis and cortical development in a mouse model of 22q11 deletion/DiGeorge syndrome. *Proceedings of the National Academy of Sciences*, *106*(38), 16434-16445.

McDonald-McGinn, D.M., Emanue,l B.S., Zackai, E.H. 22q11.2 Deletion Syndrome. (1999). In: Pagon RA, Adam MP, Ardinger HH, Wallace SE, Amemiya A, Bean LJ, et al., (Eds.), *GeneReviews* ®. Seattle, WA: University of Washington, Seattle. Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK1523/.

McDonald-McGinn, D. M., Kirschner, R., Goldmuntz, E., Sullivan, K., Eicher, P., Gerdes, M., … Zackai, E. H. (1999). The Philadelphia story: the 22q11.2 deletion: report on 250 patients. *Genetic Counseling*, *10*(1), 11–24.

Meganathan, K., Jagtap, S., Wagh, V., Winkler, J., Gaspar, J. A., Hildebrand, D., … Sachinidis, A. (2012). Identification of thalidomide-specific transcriptomics and proteomics signatures during differentiation of human embryonic stem cells. *PloS One*, *7*(8), e44228. https://doi.org/10.1371/journal.pone.0044228

Michaelovsky, E., Frisch, A., Carmel, M., Patya, M., Zarchi, O., Green, T., … Gothelf, D. (2012). Genotype-phenotype correlation in 22q11.2 deletion syndrome. *BMC Medical Genetics*, *13*(1). https://doi.org/10.1186/1471-2350-13-122

Simons VIP consortium, Green Snyder, L., D'Angelo, D., Chen, Q., Bernier, R., Goin-Kochel, R. P., … Hanson, E. (2016). Autism Spectrum Disorder, Developmental and Psychiatric Features in 16p11.2 Duplication. *Journal of Autism and Developmental Disorders*, *46*(8), 2734–2748. https://doi.org/10.1007/s10803-016-2807-4

Ousley, O., Evans, A. N., Fernandez-Carriba, S., Smearman, E. L., Rockers, K., Morrier, M. J., … Cubells, J. (2017). Examining the Overlap between Autism Spectrum

Disorder and 22q11.2 Deletion Syndrome. *International Journal of Molecular Sciences, 18*(5). https://doi.org/10.3390/ijms18051071

Paronett, E., Meechan, D., Karpinsky, B., LaMantia, A., & Maynard, T. (2014). Ranbp1, Deleted in DiGeorge/22q11.2 Deletion Syndrome, is a Microcephaly Gene That Selectively Disrupts Layer 2/3 Cortical Projection Neuron Generation. *Cerebral Cortex,* https://doi.org/10.1093/cercor/bhu285.

Rees, E., Kirov, G., Sanders, A., Walters, J. T. R., Chambert, K. D., Shi, J., … Owen, M. J. (2014). Evidence that duplications of 22q11.2 protect against schizophrenia. *Molecular Psychiatry*, *19*(1), 37–40. https://doi.org/10.1038/mp.2013.156

Rump, P., de Leeuw, N., van Essen, A. J., Verschuuren-Bemelmans, C. C., Veenstra-Knol, H. E., Swinkels, M. E. M., … van Ravenswaaij-Arts, C. M. (2014). Central 22q11.2 deletions. *American Journal of Medical Genetics Part A*, *164*(11), 2707–2723. https://doi.org/10.1002/ajmg.a.36711

Rutter, M., Bailey, A., & Lord, C. (2003). *The Social Communication Questionnaire.* Los Angeles: Western Psychological Services.

Schneider, M., Debbané, M., Bassett, A. S., Chow, E. W. C., Fung, W. L. A., van den Bree, M., … International Consortium on Brain and Behavior in 22q11.2 Deletion Syndrome. (2014). Psychiatric disorders from childhood to adulthood in 22q11.2 deletion syndrome: results from the International Consortium on Brain and Behavior in 22q11.2 Deletion Syndrome. *The American Journal of Psychiatry*, *171*(6), 627–639. https://doi.org/10.1176/appi.ajp.2013.13070864

Sparrow, S., Cicchetti, D., & Balla, D. (2005). *Vineland Adaptive Behavior Scales, 2nd edition*. London: Pearson.

Strömland, K., Nordin, V., Miller, M., Akerström, B., & Gillberg, C. (1994). Autism in thalidomide embryopathy: a population study. *Developmental Medicine and Child Neurology*, *36*(4), 351–356.

Weisfeld-Adams, J. D., Edelmann, L., Gadi, I. K., & Mehta, L. (2012). Phenotypic heterogeneity in a family with a small atypical microduplication of chromosome 22q11.2 involving TBX1. *European Journal of Medical Genetics*, *55*(12), 732–736. https://doi.org/10.1016/j.ejmg.2012.08.011

Wenger, T. L., Kao, C., McDonald-McGinn, D. M., Zackai, E. H., Bailey, A., Schultz, R. T., … Hakonarson, H. (2016). The Role of mGluR Copy Number Variation in Genetic and Environmental Forms of Syndromic Autism Spectrum Disorder. *Scientific Reports*, *6*, 19372. https://doi.org/10.1038/srep19372

Wenger, T. L., Miller, J. S., DePolo, L. M., de Marchena, A. B., Clements, C. C., Emanuel, B. S., … Schultz, R. T. (2016). 22q11.2 duplication syndrome: elevated rate of autism spectrum disorder and need for medical screening. *Molecular Autism*, *7*, 27. https://doi.org/10.1186/s13229-016-0090-z

Zhang, M. S., Arnaoutov, A., & Dasso, M. (2014). RanBP1 governs spindle assembly by defining mitotic Ran-GTP production. *Developmental Cell*, *31*(4), 393–404. https://doi.org/10.1016/j.devcel.2014.10.014

Tables

Table 1.

Descriptive characteristics of all participants in study

Region	**N**	**% de novo[a]**	**Age mean(sd)**	**Age range (years)**	**% Male**
Total	**46**	**60%**	**10.8(10.1)**	**0.8-39**	**52%**
AB/AC del group	15	86%	8.9 (4.2)	2-15	53%
A-B Deletion	11	80%	7.8(3.8)	2-15	55%
A-C Deletion	4	100%	11.8(4.4)	5-14	50%
A-B Duplication[b]	1	unknown	7.0	-	0%
BD/CD del group	18	69%	11.1(10.7)	1-38	50%
B-D Deletion	14	77%	10.4(9.0)	1-38	43%
C-D Deletion	4	33%	13.6(16.9)	0.8-36	75%
BD/CD dup group	10	13%	14.5(15.2)	1-39	60%
B-D Duplication	8	0%	16.5 (16.6)	1-39	63%
C-D Duplication	2	50%	6.5(2.0)	5-7	50%
E-F Duplication[b]	2	0%	6.6(2.8)	4-8	50%

Note. Participant characteristics for all individuals with a nested deletion or duplication of 22q11.2, including 3 case studies with atypical nested duplications.

[a] of individuals with known inheritance, [b]Case studies not included in statistical analysis, medical chart review, or AB/AC and BD/CD group totals.

Table 2.

Descriptive characteristics of participants included in psychiatric diagnosis rates

Region	**N**	**% de novo[a]**	**Age mean(sd)**	**Age range (years)**	**% Male**
AB/AC del group	13	83%	10 (4.2)	5-18	54%
A-B Deletion	10	78%	8.9 (3.6)	5-15	60%
A-C Deletion	3	100%	13.7 (4.8)	9-18	33%
BD/CD del group	12	50%	14.2 (12.7)	3-42	50%
B-D Deletion	8	57%	13.4 (12.1)	4-42	38%
C-D Deletion	4	33%	15.9 (15.5)	3-37	75%
BD/CD dup group	9	14%	16.9 (14.7)	5-39	56%
B-D Duplication	7	0%	19.4 (15.9)	5-39	57%
C-D Duplication	2	50%	8 (4.1)	5-11	50%
Total	34	55%	13.3 (11)	3-42	53%

Note. Participant characteristics for the subset of individuals with a nested deletion or duplication of 22q11.2 included in description of psychiatric diagnosis rates.

[a] of individuals with known inheritance.

Table 3.

ASD rates among probands

Region	N	n ASD (male)	% de novo [a]	Age mean(sd)	Age range (years)	% Male
AB/AC del	12	5(3)	90%	10.0 (4.4)	5-18	58%
A-B	9	4(3)	90%	8.7 (3.7)	5-15	67%
A-C	3	1(0)	100%	13.7 (4.8)	9-18	33%
BD/CD del	8	0	60%	10.5 (4.8)	5-18	25%
B-D	6	0	70%	10.2 (3.8)	6-17	17%
C-D	2	0	0%	11.6 (9.1)	5-18	50%
Classic AD dup	29	7(5)	67%	7.1 (3.4)	2-13	75%
BD/CD dup	5	1(0)	30%	12.7 (10.4)	5-31	60%
B-D	3	1(0)	0%	15.8 (13.1)	7-31	67%
C-D	2	0	50%	8.0 (4.1)	5-11	50%

Note. Participant characteristics and autism diagnosis for all probands with a nested deletion of 22q11.2. Individuals harboring a AB or AC deletion presented with ASD at 41.6%. Case studies are excluded from this table.

[a] of individuals with known inheritance.

Table 4.

Group means and effect sizes of group differences on neuropsychiatric questionnaires

	N	Mean (SD)	*d*	95% Confidence Interval
SRS-2 T score		**50 (10)**		
BD/CD deletion	9	53.4 (13.6)		
AB/AC deletion	10	68.5 (15.7)	1.01	(-0.07, 2.11)
Classic deletion	61	63 (12.6)	-0.41	(-1.11, 0.27)
BD/CD duplication	6	59.5 (16.6)		
Classic duplication	28	64 (16.3)	0.27	(-0.67, 1.22)
SCQ raw total		**cut-off: 15**		
BD/CD deletion	7	5.7 (4.9)		
AB/AC deletion	8	15 (9.3)	1.20	(-0.11, 2.52)
Classic deletion	52	10.8 (7.3)	-0.54	(-1.32, 0.23)
BD/CD duplication	4	7.2 (9.5)		
Classic duplication	22	12.4 (10.3)	0.50	(-0.67, 1.68)
Vineland Composite		**100(15)**		
BD/CD deletion	5	103.8 (19.6)		
AB/AC deletion	10	85.7 (24.0)	-0.79	(-2.11, 0.52)
Classic deletion	57	87.1 (15.9)	0.08	(-0.61, 0.78)
BD/CD duplication	6	92.6 (18.7)		
Classic duplication	27	89.4 (19.4)	-0.16	(-1.11, 0.78)

Vineland Communication		**100(15)**		
BD/CD deletion	5	106 (21.7)		
AB/AC deletion	10	83 (18.5)	-1.17	(-2.55, 0.20)
Classic deletion	58	89.7 (18.1)	0.36	(-0.32, 1.06)
BD/CD duplication	6	92.1 (20.2)		
Classic duplication	27	91.2 (18.5)	-0.04	(-0.99, 0.90)
Vineland Daily Living		**100(15)**		
BD/CD deletion	5	96.4 (14.6)		
AB/AC deletion	10	86.9 (26.2)	-0.40	(-1.69, 0.87)
Classic deletion	57	88.4 (15.1)	0.08	(-0.60, 0.78)
BD/CD duplication	6	91.8 (23.0)		
Classic duplication	28	93.6 (21.2)	0.08	(-0.86, 1.02)
Vineland Socialization		**100(15)**		
BD/CD deletion	5	107.6 (19.6)		
AB/AC deletion	10	91.0 (27.4)	-0.65	(-1.96, 0.65)
Classic deletion	57	90.1 (16.5)	-0.04	(-0.74, 0.64)
BD/CD duplication	6	98.1 (18.3)		
Classic duplication	27	90.8 (21.2)	-0.35	(-1.30, 0.60)
CASI ADHD				
BD/CD deletion	2	2.5 (1.3)		
AB/AC deletion	7	3.4 (1.4)	0.63	(-1.56, 2.83)
Classic deletion	43	2.7 (1.2)	-0.52	(-1.36, 0.32)

BD/CD duplication	4	2.7 (2.1)		
Classic duplication	20	2.8 (1.3)	0.09	(-1.09, 1.27)
CASI Anxiety				
BD/CD deletion	2	1.3 (0.7)		
AB/AC deletion	7	1.9 (1.1)	0.49	(-1.68, 2.67)
Classic deletion	43	2.3 (1.3)	0.30	(-0.53, 1.14)
BD/CD duplication	4	1.9 (1.5)		
Classic duplication	20	1.8 (1.6)	-0.05	(-1.24, 1.12)
CASI ASD				
BD/CD deletion	1	n/a		
AB/AC deletion	4	1.0 (1.0)	n/a	n/a
Classic deletion	37	0.5 (0.4)	-0.83	(-1.94, 0.27)
BD/CD duplication	4	0.6 (0.4)		
Classic duplication	16	0.8 (0.8)	0.26	(-0.98, 1.5)
CASI Schizoaffective				
BD/CD deletion	1	n/a		
AB/AC deletion	3	0.5(0.5)	n/a	n/a
Classic deletion	6	0.5(0.2)	-0.09	(-0.78, 0.61)
BD/CD duplication	0	n/a		
Classic duplication	4	0.5(0.2)	n/a	n/a
CASI Behav. Regulation				
BD/CD deletion	2	0.9 (0.4)		

AB/AC deletion	7	0.9 (0.6)	0.06	(-2.08, 2.21)
Classic deletion	43	1.0 (0.6)	0.16	(-0.67, 1.00)
BD/CD duplication	4	0.3 (0.3)		
Classic duplication	20	1.0 (0.8)	0.78	(-0.42, 2.00)
CASI Depression				
BD/CD deletion	2	0.4 (0)		
AB/AC deletion	7	0.6 (0.7)	0.33	(-1.82, 2.50)
Classic deletion	43	0.5 (0.7)	-0.15	(-0.99, 0.67)
BD/CD duplication	4	0.2 (0.4)		
Classic duplication	20	0.4 (0.4)	0.48	(-0.70, 1.68)

Note. Group means on neuropsychiatric questionnaires. We show 95% confidence intervals of effect sizes as Cohen's *d*, which can be interpreted as follows: 0.2 as small, 0.5 as medium, and 0.8 as large (Cohen, 1988). Means and standard deviations for each group are presented, as well as the mean and SD for each measure to aid in interpretation. We derived SRS T-scores using the updated SRS-2 norms for all participants, regardless of the version the participant completed. We averaged CASI-4R raw item scores on similar subscales instead of using *T*-scores because we encountered a strong ceiling effect when using CASI-4R *T*-scores because CASI-4R norms collapse all high raw scores into a *T*-score of 70, and thus population-normed means and standard deviations are not available for comparison. For example, all items from the dysthymia subscale and major depression subscales were averaged into a "Depression" composite, after accounting for the number of items in each subscale so that both scales were weighted equally in the

composite. The composites are interpreted as ‘3’ indicating that on average, the parent endorsed symptoms in the domain as occurring ‘very often,’ 2 as ‘often’, 1 as ‘sometimes,’ and 0 as ‘never.’ Vineland = Vineland Adaptive Behavior Scales, 2nd Edition; CASI = Child and Adolescent Symptom Inventory-4R; SCQ = Social Communication Questionnaire; SRS-2 = Social Responsiveness Scale, 2nd Edition

Table 5.

Psychiatric disorder rates from parent and adult self-report and chart review

	Total (n)	No Eval	Had Eval	GDD	Lang Dx	ADHD	ID	ODD	OCD	Anxiety	MDD
AB/AC Del	13	7.7%	92.3%	30.8%	0.0%	53.8%	0.0%	7.7%	23.1%	23.1%	0.0%
A-B Deletion	10	10%	90%	30%	0%	50%	0%	10%	20%	20%	0%
<3 yrs	1	1	0	0	0	0	0	0	0	0	0
3-14yrs	8	0	8	2	0	4	0	0	1	1	0
15+yrs	1	0	1	1	0	1	0	1	1	1	0
A-C Deletion	3	0%	100%	33.3%	0%	66.7%	0%	0%	33.3%	33.3%	0%
3-14yrs	1	0	1	1	0	0	0	0	0	0	0
15+yrs	2	0	2	0	0	2	0	0	1	1	0
BD/CD Del	12	4	8	2	1	1	3	1	3	3	0
B-D Deletion	8	12.5%	87.5%	37.5%	0%	12.5%	12.5%	0%	0%	25%	25%

<3 yrs	1	1	0	0	0	0	0	0	0	0	0
3-14yrs	5	0	5	3	0	1	1	0	0	0	0
15+yrs	2	0	2	0	0	0	0	0	0	2	2
C-D Deletion	4	75%	25%	0%	0%	25%	0%	25%	25%	25%	25%
<3 yrs	2	2	0	0	0	0	0	0	0	0	0
15+yrs	2	1	1	0	0	1	0	1	1	1	1
BD/CD Dup	9	3	6	2	0	0	4	0	1	1	1
B-D	7	42.9%	57.1%	28.6%	14.3%	28.6%	0%	0%	0%	14.3%	14.3%
Duplication											
<3 yrs	1	1	0	0	0	0	0	0	0	0	0
3-14yrs	3	0	3	2	1	1	0	0	0	1	0
15+yrs	3	2	1	0	0	1	0	0	0	0	1
C-D											
Duplication											

3-14yrs	2	0	2	2	0	0	0	0	0	0	0
Total Sample	**34**	**23.5%**	**76.5%**	**32.3%**	**2.9%**	**32.3%**	**2.9%**	**5.9%**	**11.8%**	**20.6%**	**11.8%**

Note. We observed elevated rates of psychiatric diagnoses among individuals with nested duplications or deletions relative to population base rates using parent- and self- report data confirmed in medical records. Among the full sample, 77% had received a psychiatric evaluation. The most commonly reported diagnoses in our sample included ADHD and Global Developmental Delay (GDD), which may reflect the sample's skew toward younger ages (see table 2 for sample characteristics). We present rates for group totals, and we present n's for age bins based roughly on when documentation of diagnosis would be expected (*i.e.*, GDD and language disorders are frequently diagnosed before age 3, ADHD and ID are usually diagnosed in childhood after age 3, and depression and anxiety frequently onset during adolescence or adulthood) to facilitate interpretation of overall group rates because rates for disorders that frequently appear in adolescence (e.g., anxiety and depression) are likely underestimates. Abbreviations: Eval =evaluation; ADHD =Attention Deficit/Hyperactivity Disorder; ID = Intellectual Disability; OCD =Obsessive Compulsive Disorder; GDD = Global Developmental Delay; Lang Dx = Language Disorder, receptive or expressive; ODD = Oppositional Defiant Disorder; MDD = Major Depressive Disorder; Del = Deletion; Dup = Duplication;yrs= years

Table 6.

Medial comorbidities in individuals with nested deletions and duplication of 22q11.2

	A to B deletion (n=11)	**A to C deletion (n=4)**	**B to D deletion (n=14)**	**B to D duplication (n=8)**	**C to D deletion (n=4)**	**C to D duplication (n=2)**
Audiologic						
Audiogram	11	4	14	8	4	2
Abnormal	6	1	2	2	0	1
Abnormal Rate[a]	55% CHL	25% CHL	14% SNHL	25% CHL, CSNHL	0% n/a	50% CHL
Abnormalities						
Cardiac						
Echocardiogram	9 7	4 3	11 7	5 1	2 1	1 0
Abnormal	64% (78%)	75%	50% (64%)	12% (20%)	25% (50%)	0%
Abnormal Rate Abnormalities	Enlarged PA, VR, PS, TOF with PS, IAA with ARSCA, TR, PDA	TA, RAA with ALSCA, dilated aortic root, VR, ASD/VSD	Aortic root dilation, aneurism of TV, ASD, PDA, PFO, TA, VSD	PFO	TOF with Pulmonary valve stenosis	n/a
Endocrine						
Bloodwork	11	4	14	4	4	1
Abnormal	6	2	6	2	0	0
Abnormal	55% Hypocalcemia	50%	43% Hypocalcemia	25% (50%) Borderline	0%	0%

Rate Abnormalities	hypothyroidism low vitamin D (Each category n=2)	Hypocalcemia	(n=2), diabetes mellitus, borderline HbA1C, low growth factors, low vitamin D, hypothyroidism	abnormal thyroid function tests, neonatal hypoglycemia	n/a	n/a
GI						
Symptom	11	3	13	6	3	2
screen	7	2	12	5	1	2
Abnormal	64%	50% (67%)	86% (92%)	63% (83%)	25% (33%)	100%
Abnormal Rate Abnormalities	GERD (n=7), constipation (n=4), anal atresia (n=1), feeding tube (n=2)	Constipation (n=2), GERD	GERD (n=11), constipation (n=9), feeding tube (n=4)	GERD (n=4), eosinophilic esophagitis, feeding tube	GERD, chronic constipation	GERD (n=2), constipation, feeding tube
Hematologic						
CBC	11	4	14	8	4	2
completed	3	3	2	0	0	0
Cytopenias	27%	75%	14%	0%	0%	0%
Abnormal Rate						
Immune						
Bloodwork	9	4	11	4	4	2
Abnormal	2	2	4	1	2	0
Abnormal	18% (22%)	50%	29% (36%)	13% (26%)	50%	0%

Rate Abnormalities	Low Ig	Low Ig, T-cell lymphopenia, inadequate vaccine titers	Low Ig (n=2), absent thymus, inadequate vaccine response	Low Ig	Low Ig, recurrent MRSA infections, inadequate vaccine response	n/a
Neurologic						
Seizures	3	0	2	0	0	0
% Reported	27%	0%	14%	0%	0%	0%
MRI	8	2	10	4	0	0
Abnormal	3	1	3	2	n/a	n/a
MRI Abnormal Rate	27% (38%)	25% (50%)	21% (30%)	25% (50%)		
MRI Findings	Chiari 1, white matter lesions, pachygyria	Minimal bilateral congenital optic nerve hypoplasia	Chiari 1, hypoplastic corpus callosum, polymicrogyria	Prominent ventricles, subarachnoid spaces, choroid plexus cysts		
Ophthalmology	9	4	11	3	4	1
Ophtho exam	3	2	5	1	0	0
Abnormal	27% (33%)	50%	45%	13% (33%)	n/a	n/a
Abnormal Rate	Astigmatism, exophoria, nystagmus	Strabismus, minimal ONH	Anisocoria, iris coloboma, ONH nystagmus (n=2), retinal detachment, strabismus (n=3)	Amblyopia		
Abnormalities						

Palate						
Clinical eval.	9	4	12	6	2	2
Abnormal	8	3	4	1	0	1
Abnormal Rate	89% (73%)	75% (27%)	33% (29%)	17% (13%)	0%	50%
Abnormalities	SMCP (n=3) VPI (n=8)	SMCP (n=1), VPI (n=3)	SMCP (n=2), VPI (n=4)	High arched palate with small uvula	n/a	VPI
Renal						
Ultrasound	9	2	9	6	2	1
Abnormal	3	0	1	4	1	0
Abnormal Rate	27% (33%)	0%	7% (11%)	50% (67%)	25% (50%	0%
Abnormalities	Bilateral pelviectasis, nephrocalcinosis hydronephrosis	n/a	Medullary nephrocalcinosis	Duplicated collecting system, small kidneys (n=3)	Solitary, low-lying kidney	n/a
Spine						
Screening x-rays	6	4	4	2	1	1
Abnormal	6	3	2	1	1	0
Abnormal Rate	55% (100%)	38% (75%)	14% (50%)	13% (50%)	25% (100%)	0%
Abnormalities	Hypoplastic vertebra (n=2), vertebral fusion (n=4), extra lumbar vertebra	Fusion of C2-C3, kyphoscoliosis, thickened spinous process of C2	Scoliosis, C2-C3 fusion and dysmorphic dens, upswept C2	Hemivertebra at T9, absent rib	6 thoracic ribs and 6 lumbar vertebrae	n/a

*Note***.** The total number of patients in each group is designated in column headings. Each screened organ system is listed along with the number of patients who received the screening recommended for patients with classic 22q11.2 deletions and duplications. We present the patients with abnormal findings as percentage of total patients. Many patients did not receive all recommended screening; when not all patients were screened, we use parentheses to note the percentage of patients with abnormal findings among those who received screening. Abbreviations**:** ARSCA Aberrant right subclavian artery; ALSCA Aberrant left subclavian artery; ASD Atrial septal defect (in Cardiac row only; in remainder of manuscript ASD refers to autism spectrum disorder); CHL Conductive hearing loss; C/SNHL Mixed conductive and sensorineural hearing loss; GERD Gastroesophageal reflux disease; HbA1C Hemoglobin A1C; IAA Interrupted aortic arch; Ig Immunoglobulins; MRI Magnetic resonance imaging; MRSA Methicillin-resistant Staphylococcus aureus; ONH Optic nerve hypoplasia; PA Pulmonary artery; PDA Patent ductus arteriosus; PFO Patent foramen ovale; PS Pulmonic stenosis; SMCP Submucous cleft palate; SNHL Sensorineural hearing loss; TA Truncus arteriosus; TR Tricuspid regurgitation; TOF Tetralogy of Fallot; TV Tricuspid valve; VPI Velopharyngeal insufficiency; VR Vascular ring; VSD Ventricular septal defect

Figures

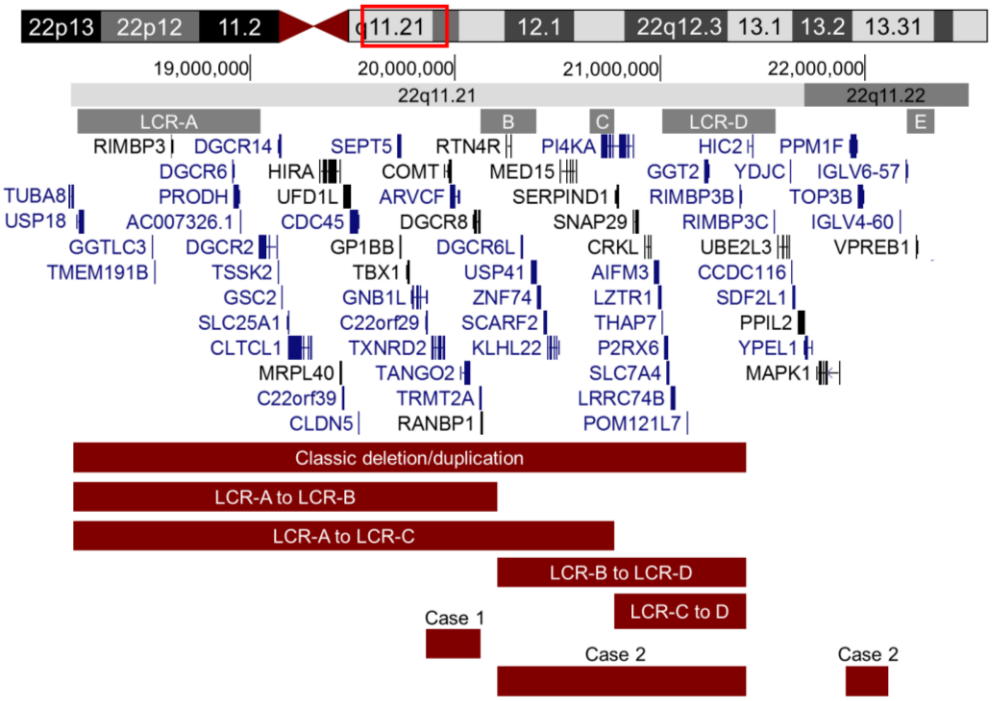

Figure 1. 22q11.2 diagram. Genes and Low Copy Repeat ("LCR") regions in the 22q11.2 region. Red bars depict deletions or duplications of participants. From GENCODE v24 genes in UCSC genome browser, December 2013 Assembly (genome.ucsc.edu)

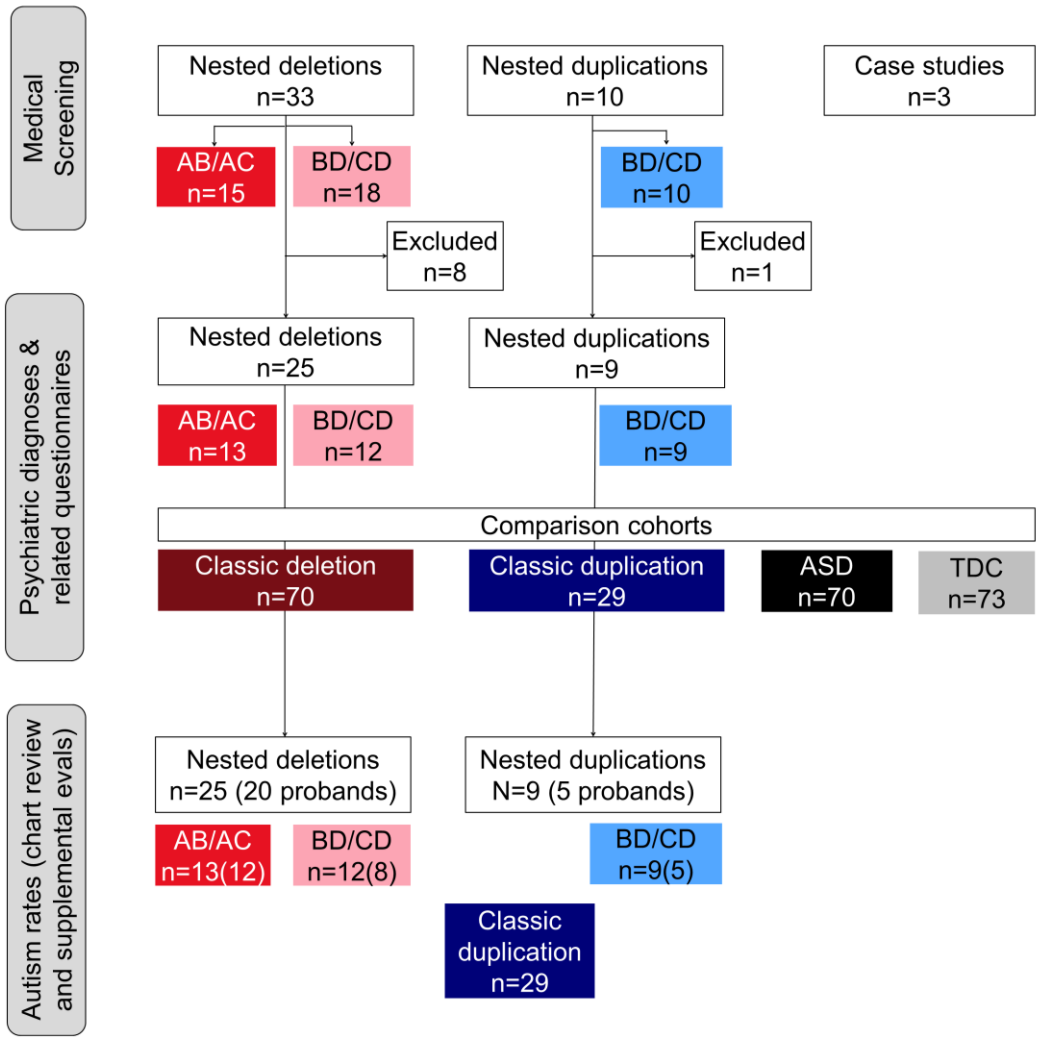

Figure 2. Participant flow chart. The participants and comparisons included in each portion of the study. Group colors correspond to colors in figures 3, 4, and additional file 2. Abbreviations: AB/AC: deletion or duplication spanning LCR-A to LCR-B, or LCR-A to LCR-C; ASD: autism spectrum disorder; BD/CD: deletion or duplication spanning LCR-B to LCR-D, or LCR-C to LCR-D; d; Cohen's *d* effect size; del: typical 22q11.2 Deletion Syndrome involving LCR-A to D, dup: typical 22q11.2 Duplication Syndrome involving LCR-A to D; LCR: Low Copy Repeat region; TDC: typically developing

controls

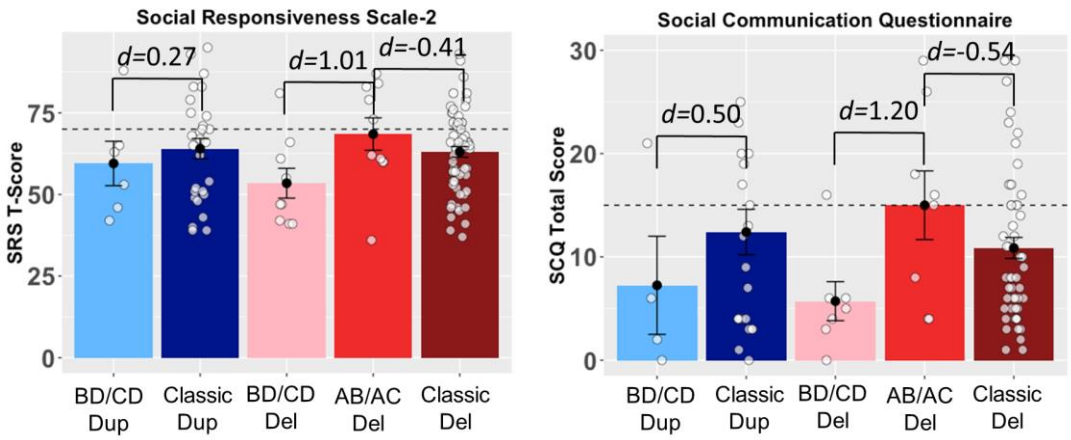

Figure 3. Individuals with deleted LCR-A to B show higher levels of autistic symptoms. Group means, standard errors, and the effect size of differences on two quantitative measures of autistic symptoms, the Social Responsiveness Scale-2 (SRS-2) and the Social Communication Scale, Lifetime (SCQ). Each point depicts one individual. Dashed lines indicate the threshold above which an individual is considered to screen positive for autism and warrant further evaluation. The "BD/CD dup" (light blue) and "BD/CD del" (light pink) groups include individuals with duplications or deletions, respectively, of LCR-B to D or LCR-C to D. The comparison groups include individuals with duplicated or deleted LCR-A to B; for duplications, the "Classic Dup" group (dark blue) includes individuals with the classic duplication of LCR-A to D, and for deletions, the "AB/AC del" group (red) includes individuals with nested deletions of LCR-A to B or C while the "Classic Del" group (dark red) includes individuals with classic deletion of LCR-A to D. The groups with involvement of LCR-A to B show higher levels of social impairment, with large effect sizes for deletions and small to medium effect sizes for duplications. Effect sizes are not significant due to small samples (see table 4). The AB/AC deletion

group includes 5 individuals diagnosed with autism; the BD/CD deletion group includes zero. Abbreviations: AB/AC: deletion or duplication spanning LCR-A to LCR-B, or LCR-A to LCR-C; ASD: autism spectrum disorder; BD/CD: deletion or duplication spanning LCR-B to LCR-D, or LCR-C to LCR-D; d; Cohen's *d* effect size; del: typical 22q11.2 Deletion Syndrome involving LCR-A to D, dup: typical 22q11.2 Duplication Syndrome involving LCR-A to D; LCR: Low Copy Repeat region; TDC: typically developing controls; SCQ: Social Communication Questionnaire, Lifetime; SRS: Social Responsiveness Scale

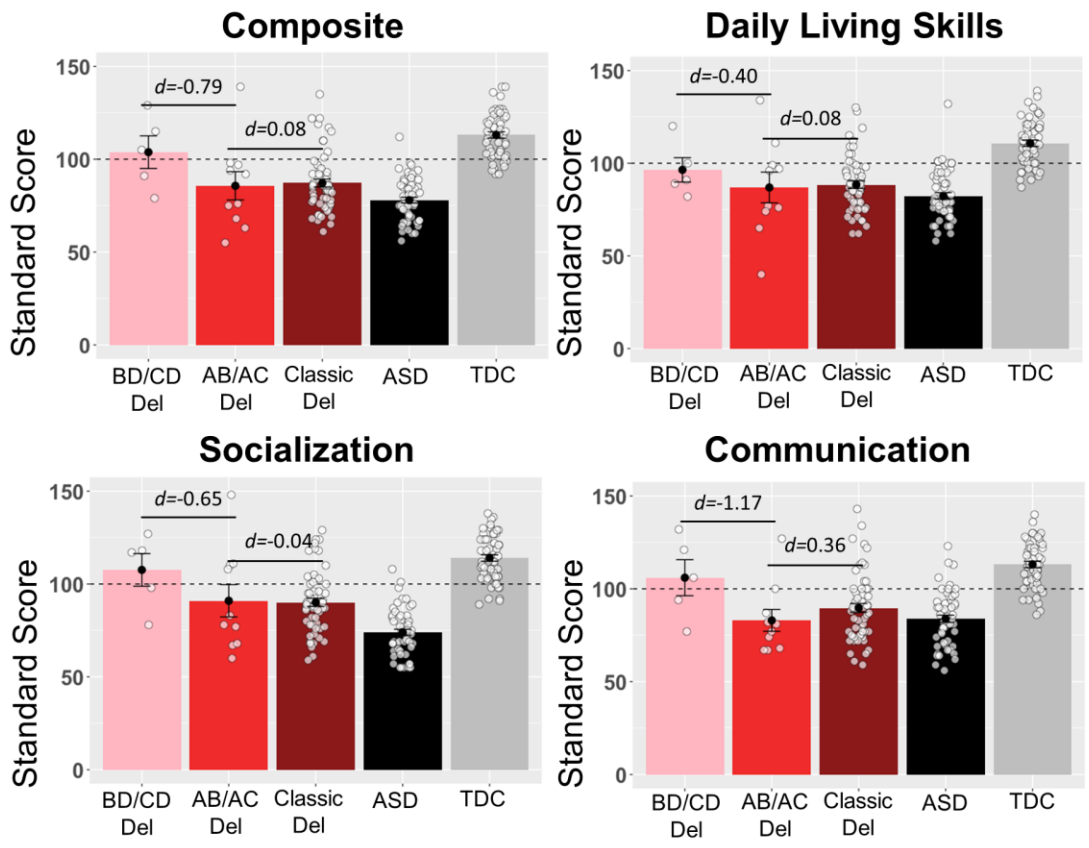

Figure 4. Individuals with deleted LCR-A to B show modestly lower levels of adaptive functioning on the Vineland-II Adaptive Behavior Scales. Group means, standard errors, and the effect size of differences on the Vineland-II, a measure of adaptive behavior. Each point depicts one individual. Groups include the "BD/CD" deletion group (light pink; individuals with nested duplication or deletion involving LCR-B or C to LCR-D), the "AB/AC" deletion group in red (individuals with nested deletion of LCR-A to B or C), the "Classic Del" group in bar red (individuals with typical deletion spanning LCR-A to LCR-D), "ASD" group in black (individuals with non-syndromic autism spectrum disorder), and "TDC" group in green (typically developing children). Higher scores on the Vineland-II indicate higher levels of functioning across the three domains and

composite score, and standard scores are age-normed such that 100 (represented by the dashed line) indicates average. The “AB/AC” deletion group shows more impairment than the “BD/CD” nested deletions that do not involve LCR-A to B with medium to large effect sizes; the “AB/AC” group also shows similar levels of impairment to the classic delection group, with small or negligible effect sizes. Abbreviations: AB/AC: deletion spanning LCR-A to LCR-B, or LCR-A to LCR-C; ASD: autism spectrum disorder; BD/CD: deletion spanning LCR-B to LCR-D, or LCR-C to LCR-D; d; Cohen’s d effect size; del: typical 22q11.2 Deletion Syndrome involving LCR-A to D, dup: typical 22q11.2 Duplication Syndrome involving LCR-A to D; LCR: Low Copy Repeat region; TDC: typically developing controls

CHAPTER 2: FEASIBILITY OF SMALL SAMPLES TO DEVELOP A LARGE ITEM POOL FOR COMPUTER ADAPTIVE TESTING, WITH EMPIRICAL AND SIMULATED DATA

Abstract

The use of small samples to develop computer adaptive test (CAT) item pools would make developing a CAT assessment feasible in small, non-commercial settings. This simulation study investigates the possibility of using small samples (N=300, 500) to calibrate items for a large CAT item pool. The study answers this question in the context of developing an IQ CAT with multiple groups of different mean IQs (n=8 groups) for use in psychiatric genetic research, and leverages empirically derived item parameters from a previous study. Two study factors were manipulated to examine their effect on linking error (the error introduced by linking data across groups): the proportion of common items across groups (one-third and one-half), and the type of parameter calibration (concurrent vs. separate calibration followed by Stocking-Lord linking). Linking error in optimal conditions was compared between small and more traditional sample sizes (e.g., N=1000, N=3000). Results indicated that at small samples sizes, concurrent calibration resulted in significantly less linking error than Stocking-Lord linking, and the proportion of common items used in linking had no appreciable effect. However, the N=300 condition resulted in a significantly higher proportion of model fit issues, making N=300 per group a risky sample size. Although the choice of study factors will depend on cost-benefit analysis and the tolerance for error of individual developers, in the present example, one particular condition (N=500, concurrent calibration, 20 common items) proved superior with regard to model failures, linking error in item and ability parameters, and total size number of item pool.

Introduction

A research subject's general cognitive ability is frequently captured by measuring IQ. IQ is arguably one of the most extensively studied and best understood constructs in all of psychology, with origins dating back to the turn of the 20th century, when it was frequently noted that scores on various cognitive ability tests usually correlated positively with one another (Spearman, 1904). The single common factor underlying these correlations was defined as *g*, the general factor of intelligence. Intelligence predicts or interacts with many constructs of interest, and is frequently assessed by IQ tests in both research and clinical practice, particularly in pediatric settings. Twin and family genetic studies long ago demonstrated that IQ is heritable (for review, see Knopik, Neiderhiser, DeFries, & Plomin, 2016), and modern methods (e.g., SNP arrays and genome-wide association studies (GWAS)) have identified common genetic variants associated with IQ (e.g., Zabaneh et al., 2017; Savage et al., 2018; for review, see Plomin & von Stumm, 2018). The genetic basis of IQ, however, is a highly controversial topic in the media; researchers continue to study this important construct with new genomic methods ("Intelligence Research Should Not Be Held Back by its Past," 2017).

The lack of a reliable, brief, online, and inexpensive IQ assessment severely limits modern psychiatric genomic research, where assessment development resources are scarce. To the detriment of psychiatric genomic research, it is infeasible to assess very large samples using reliable, gold-standard IQ tests such as the Wechsler Abbreviated Scales of Intelligence (Wechsler, 2011). Instead, the effort to conduct GWAS on cognitive ability to date has relied heavily on the proxy variable of educational attainment

(Rietveld et al., 2014; Davies et al., 2016; Lee et al., 2018), which is only weakly correlated with IQ (Calvin et al., 2012) and introduces socioeconomic confounds (Bates, Lewis, and Weiss, 2013; Braveman et al., 2005). Accordingly, many have cited the need for a brief measure of IQ (*e.g.*, Krasileva, Sanders, & Hus Bal, 2017). Computer Adaptive Testing (CAT) is particularly useful to this end as it reduces the number of items administered by 50% or more (Weiss, 2004), offers better precision than fixed length tests, particularly at the high and low ends of the ability range (Wainer, Dorans, Flaugher, Green, & Mislevy, 2000), and increases self-reported motivation in low ability test takers (Betz, 1977). A computer-adaptive, online IQ assessment could be widely administered and, if carefully developed and validated, offer a substantial improvement in reliability and validity over educational attainment and other metrics with unestablished psychometric properties currently in use.

The authors attempted to develop such a tool, and found that most relevant methodological literature relies on large samples that are untenable for assessment development by a small, non-commerical research group. The sample size issue is exacerbated when the assessment requires multiple nonequivalent groups, or groups with different mean abilities. The proposed psychiatric genetic IQ assessment requires data collection from nonequivalent groups because cognitive ability changes with age (Deary et al., 2009; Hedden & Gabrieli, 2004) and ages are usually binned for IQ scoring (Wechsler, 2011). Many existing studies of CATs for nonequivalent groups leverage NEAT designs (NonEquivalent groups with Anchor Test), referring to patchwork designs in which every group receives a subset of items, with some common items across groups,

and one group is chosen as the reference group or anchor test (see Figure 1 for example). Many existing CAT studies also use large educational datasets (e.g., Measures of Academic Progress: Wang, McCall, Jiao, & Harris, 2012) and require sample sizes that are prohibitively large for small resource settings. This study explores potential solutions to the challenges encountered so that small research groups can develop CATs that reap the benefits of large item pools, including for assessing IQ for psychiatric genetics research.

Challenges of developing a CAT item bank with minimal resources

The proposed assessment will feed questions from a calibrated item pool into the CAT of a verbal IQ assessment. The verbal IQ assessment consists of multiple choice vocabulary items, since vocabulary subtests of traditional IQ tests show high factor loadings on the general intelligence factor *g* (λ=0.74, Wechsler, 2008). A large, calibrated CAT item pool requires multiple samples of examinees to respond to subsets of items, since no single examinee can sustain attention to respond to all items in the target bank size of 300 items. Smaller item sets administered to independent examinee samples are pieced together to form the total item bank. CAT item bank developers then face the challenge of linking, or translating all item parameters onto a common scale. This problem arises because each group of examinees that completes responses for a set of items has a unique ability distribution, but each calibration typically assumes the same normal distribution N(0,1), rendering not only ability but also item parameter estimates from each sample incomparable because they are on different scales. All items in the final pool require one common scale. To link the item sets and samples, several ‘common

items' are administered across groups. Then, one group is chosen as the reference group with a normal distribution N(0,1) of estimated abilities, and other group parameters are transformed onto this single common scale using items common across all groups. Each of these design decision points and computational steps introduces bias, which different study conditions may disproportionately magnify in the small sample sizes of a non-commerical setting.

Feasibility of very small sample sizes

Much prior research describes how larger sample sizes improve model fit and decrease error in item parameters (discrimination, difficulty, and guessing) and examinee ability parameters (θ). Lord (1968) reported difficulties with model convergence using SAT data and a single group, and recommended studies use a minimum of 1000 examinees. However, the common wisdom that fitting a three-parameter model requires 1,000 (Lord, 1968) or 1,500 (Kolen & Brennan, 2010) examinees is prohibitive for many non-commerical research endeavors. Few studies exist on smaller samples (e.g., 300 per group), but such small samples could make developing a CAT accessible to smaller research groups. It may be possible to offset the increased bias of very small sample sizes by optimizing other study factors such as the proportion of common items, the NEAT design, and the calibration method. This study investigates whether small sample sizes (e.g., 300 per group) may be tenable if other study factors are optimized to minimize bias.

Optimizing the proportion of common items

One study factor that could be leveraged to minimize bias is the proportion of common items. During item bank calibration, each sample of examinees responds to a

fixed number of items, and a set of common items appears on all tests to facilitate linking onto a single scale. The relationship between the proportion of common items and parameter bias has been the subject of much study. Hanson and Beguin (2002) found linking error decreased when the number of common items was increased from 10 to 20 on a 60-item test administered to nonequivalent groups, although a sample size increase from 1,000 to 3,000 showed a much larger effect on decreasing bias. Kang and Petersen (2012) reported in a simulation study of a 50-item test with 10, 20, or 40 linking items that linking performance improved with higher proportions of common items, but 20% may be sufficient to obtain tolerable amounts of error. Arai and Mayekawa (2011) tested 4, 8, and 12 common items in a 40-item test length: the 12 common item condition performed best, but the extent of the benefit depended on other study factors (namely, design and calibration). Of note, they concluded from their sample of nonequivalent groups—with mean differences similar to those in the present study—that linking error was minimized when common items were shared by all groups. Thus to minimize linking error in the present study, we presented the same set of common items to all groups, instead of presenting overlapping items to adjacent groups in a patchwork fashion.

Higher proportions of common items result in smaller total item banks because examinees can only sustain attention for a fixed number of items and limited resources often dictate the total possible number of examinees. Thus, CAT bank developers must balance the decreased bias from a higher proportion of common items with the cost of additional examinees to calibrate additional items to fill the bank.

Optimizing the calibration method

There are two primary methods to establish a common scale for different item sets administered to nonequivalent groups. In the first method, concurrent calibration, all groups are calibrated simultaneously, with one group being specified as a reference group. In the second method, all groups are calibrated separately, then each group's parameters are linearly transformed onto a common scale. Linear transformation constants can be obtained via mean/sigma (Marco, 1977), mean/mean (Loyd & Hoover, 1980), and characteristic curve methods (Kolen & Brennan, 2014). The Stocking Lord characteristic curve method (Stocking & Lord, 1983) has been demonstrated to yield more accurate estimates than alternative methods across a variety of conditions similar to those in the present study (Hanson & Beguin, 2002; Kim & Kolen, 2006), so it was chosen for comparison to concurrent calibration.

Other study factor conditions such as sample size and proportion of common items affect whether concurrent or separate calibration yields smaller linking error. Kim and Cohen (1998) reported that concurrent and separate calibration yielded similar results unless the study design had a small proportion of common items. Hanson and Beguin (2002) used ACT Math subtest data and reported that the unique items for the non-reference group had higher linking error for separate calibration, compared to concurrent. They observed improvement by increasing sample size (from 1,000 to 3,000), but not by increasing the proportion of common items (from 10/60 to 20/60). Kim and Kolen (2006) also found that concurrent calibration produced lower error, closely followed by Stocking Lord, then by other linking methods such as the Haebara (1980) characteristic curve method, mean/mean, and mean/sigma methods. This study relied on simulated data based

on empirical data from a large educational test, with a sample size of 2000 examinees. It remains unknown how calibration method interacts with proportions of common items in small sample sizes.

Present study

The current study investigates whether small samples could be used to develop a large CAT item bank with nonequivalent group data by minimizing linking error with other study factors. This question is particularly relevant for developers in minimal resource settings as small samples could facilitate and expedite development of assessment tools, such as an IQ assessment, which is desperately needed for psychiatric genetic research. Using empirical parameters and simulated response data, this study compares concurrent and separate group calibration with different sample sizes and different proportions of common items to identify the set of conditions that would minimize linking error and thereby optimize the accuracy of a nonequivalent group assessment developed with the smaller sample sizes available to most researchers. The study factors evaluated are:

1. Sample size (N=300, 500, 1000, 3000), as smaller samples decrease cost
2. Proportion of common items (1/3 and 1/2), as administering fewer common items per group allows more unique items to be developed for the same cost (but may increase linking error
3. Concurrent or separate calibration followed by Stocking Lord linking

Methods

Simulation study design

Sample size. Four sample sizes were examined: 300, 500, 1000, and 3000 examinees per group. Samples of 300 and 500 were selected to assess the feasibility of calibrating parameters with small samples. Samples of 1000 and 3000 were selected as control conditions to compare to the small samples and for comparability with existing literature (e.g., Hanson & Benguin, 2002).

Proportion of common items. Two proportions of common items were evaluated in a nonequivalent group common item design: 20 and 30 common items in a test length of 60 items (Figure 1). The test length administered to each examinee was maintained constant at 60 items in both conditions. Consequently, total item bank size varied across conditions such that the 30/60 common item condition had 270 total items (30 common, and 8x30 unique to each group) while the 20/30 common item condition had 340 total items (20 common and 8x40 unique to each group).

Calibration. Two procedures were compared for linking item parameters and vertically scaling scores from nonequivalent groups onto a single scale: concurrent calibration implemented with BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003), and separate calibration implemented with BILOG-MG followed by Stocking-Lord linking implemented with STUIRT (Kim & Kolen, 2004).

Data generation

Nonequivalent group distributions. Eight nonequivalent groups were created to approximate eight age groups from 18-70 with different means and standard deviations (Table 1) to reflect the growth and decay of verbal intelligence across these age groups. Age-group design was loosely based on age groups used for scoring in the WASI-II

(Wechsler, 2011). Group means and standard deviations were derived from the WASI-II vocabulary T-scores for each age group.

Examinee data. Using WinGen (Han, 2007), 200 samples of 3000 examinees for each age group were randomly drawn from each distribution (100 primary samples, 100 back-up samples). The first 300, 500, 1000, and 3000 examinees from each group were used for each respective sample size condition. One response pattern was generated for each set of examinees using WinGen. Thus, a different, independent sample and its corresponding response pattern were used for each replication, to facilitate generalization of results over different future sample data.

Item parameter data. Parameters from empirical data were used. The response dataset contained online responses from anonymous individuals to subsets of over 500 multiple choice vocabulary items developed in consultation with a linguist. Response data were analyzed with concurrent calibration with quadrature points and strong priors to promote convergence. The same item parameter set was used for every replication, as the focus of this study was to develop a CAT item bank.

Study endpoints

In total, this study presents 16 conditions (4 sample sizes x 2 common item proportions x 2 vertical scaling methods). In the concurrent calibration condition, we fit eight total models, one for each condition; in the separate calibration condition, we fit 64 models, one for each of eight age groups in each condition. We analyzed one hundred replicates of each condition, which involved dropping and replacing a dataset if one of

the primary 100 datasets failed to converge or calibrate an item. All endpoint analyses were implemented in R v3.5.2 (R Core Team, 2018).

Number of dropped replications. Replications were dropped and replaced for three reasons: the model failed to converge; an item could not be calibrated; or the standard error of an item parameter was not reported. We recorded the number of dropped replications out of all 200 runs for each condition, and counted a replication as dropped if it failed for any model in the condition. For the final analysis, any replication that was dropped from any condition was excluded from all conditions so that the same 100 datasets were used for all conditions to facilitate comparison.

Theta recovery. Each condition was evaluated on four metrics of theta recovery. First, absolute theta bias was computed for each examinee as the *absolute* difference between the true theta and the estimated theta (Equation 1). Where *i* represents an examinee and R=100 replications

Equation 1

$$absolute\ bias_i = \frac{\sum_{r=1}^{R}(|\hat{\theta}_{ir} - \theta_i|)}{R}$$

Second, signed theta bias was computed as the difference between the true theta and the estimated theta (equation 2). Where *i* represents an examinee and R=100 replications

Equation 2

$$signed\ bias_i = \frac{\sum_{r=1}^{R}(\hat{\theta}_{ir} - \theta_i)}{R}$$

Third, the standard error (SE) was computed for each examinee (Equation 3). Where *i* represents an examinee and R=100 replications

Equation 3

$$SE_i = \sqrt{\frac{\sum_{r=1}^{R}(\hat{\theta}_{ir} - \bar{\hat{\theta}}_i)^2}{R}},$$

where $\bar{\hat{\theta}}_i = \frac{\sum_{r=1}^{R}\hat{\theta}_{ir}}{R}$.

Fourth, the Root Mean Square Error (RMSE) was computed for each examinee (Equation 4). Where *i* represents an examinee and R=100 replications

Equation 4

$$RMSE_i = \sqrt{\frac{\sum_{r=1}^{R}(\hat{\theta}_{ir} - \theta_i)^2}{R}} = \sqrt{bias_i^2 + SE_i^2}$$

Each endpoint was evaluated for all examinees, and by group. Each endpoint was averaged across all examinees (N=300, 500, 1000, or 3000) to obtain mean absolute bias, signed bias, SE, or RMSE for the condition.

Item parameter recovery. Each condition was evaluated on absolute and signed mean conditional bias in threshold (*b*), slope (*a*), and guessing parameters (*c*) (Equations 5-6). Where *i* represents an item and R=100 replications, conditional slope absolute bias for one item can be represented as:

Equation 5

$$absolute\ bias_i = \frac{\sum_{r=1}^{R}(|\hat{a}_{ir} - a_i|)}{R}$$

and conditional slope signed bias for one item can be represented as:

Equation 6

$$signed\ bias_i = \frac{\sum_{r=1}^{R}(\hat{a}_{ir} - a_i)}{R}$$

Conditional absolute or signed bias was then averaged across all items (N=270 or 340) to obtain mean absolute bias or mean signed bias for the condition.

Statistics

Study endpoints were compared across conditions using ANOVA with the conservative Bonferroni correction for multiple comparisons (8 or 16) within each study factor (sample size, calibration type, and proportion of common items). Instead of ANOVA, a chi square test was used to compare proportions of dropped replicates between conditions.

Results

Theta and item parameter recovery

Sample size. As expected, absolute bias in estimates of both ability and item parameters decreased as sample size increased (Tables 2, 3 and Figures 2, 3). However, bias did not improve linearly with sample size; rather, the improvement between N=300 and N=500 conditions was similar or larger than the improvement between N=500 and N=1000, despite the latter representing a much more costly increase in sample size.

Proportion of common items. Increasing common items from 20 to 30 produced no significant improvement in absolute theta bias within each sample size of the

concurrent conditions, after correction for multiple comparisons (*p*'s 0.009-0.508; Figure 2a). In the separate calibration condition, no differences between 20 and 30 common item levels were observed in absolute theta bias for sample sizes of 300 or 500, but sample sizes of 1000 ($t(99)=3.23$, $p=0.002$) and 3000 ($t(99)=4.09$, $p<0.001$) showed significant improvements in absolute theta bias for the 30 common item level. The separate, 30 common item condition outperformed all other conditions with regard to absolute theta bias, but also showed the highest standard error.

Item parameter recovery. Absolute bias in discrimination and item difficulty parameters was nearly identical for both common item levels within each sample size and within each calibration type (all *p*'s > 0.10; Figure 3c and 3d).

Calibration type. With regard to absolute theta bias, separate calibration significantly outperformed concurrent calibration at all sample sizes except n=300 (300: $F(3,396)=0.69$, $p=0.56$; 500: $F(3, 396)=5.9$, $p=0.001$; 1000: $F(3, 396)=12.4$, $p<0.001$; 3000: $F(3, 396)=20.82$, $p<0.001$). In contrast, concurrent calibration showed significantly less signed theta bias at all sample sizes (all *p*'s<0.001). Comparisons on RMSE, which accounts for both bias and SE, demonstrated superiority of the concurrent condition for sample sizes of 300 and 500, but no differences between concurrent and separate conditions with 30 common items for sample sizes of 1000 and 3000.

Item parameter recovery. Absolute bias in discrimination and difficulty parameters showed negligible differences between calibration types (*p*'s>0.13; Figure 3c and 3d). Common items showed significantly less absolute bias in discrimination and difficulty parameters with concurrent calibration for all conditions (*p*'s<.01) in all cases

except difficulty parameters with N=3000 (p=0.0340; Figure 3a and 3b; Table 3). The concurrent condition also demonstrated superior performance with regard to signed bias (Figure 3e and 3f; Table 3). Separate calibration followed by Stocking-Lord linking showed larger underestimation than concurrent calibration for both difficulty and discrimination parameters with samples N=300 (difficulty: $F(3,1216)=7.76$, $p<0.001$; discrimination: $F(3,1216)=28.59$, $p<0.001$) and N=500 (difficulty: $F(3,1216)=6.96$, $p<0.001$; discrimination: $F(3,1216)=14.61$, $p<0.001$), and for discrimination for N=1000 (difficulty: $F(3,1216)=2.81$, $p=0.038$; discrimination: $F(3,1216)=4.7$, $p<0.003$). Signed bias showed no differences across conditions at the 3000 sample size (difficulty: $F(3,1216)=0.38$, $p=0.768$; discrimination: $F(3,1216)=1.2$, $p=0.309$). The concurrent conditions showed almost no signed difficulty bias.

Differences by group and ability level. Across all conditions the youngest group (group 18-19 years, distribution N -0.83, 0.86) showed substantially higher theta bias than all other groups (Table 2, Fig 4). In both calibration conditions, older groups showed larger absolute theta bias and younger groups showed smaller absolute bias.

Across the ability range, absolute theta bias performed as expected in every condition. Bias remained low and near its minimum for approximately two standard deviations outside the mean (Figure 5), and increased toward the tail ends of the distribution where there were fewer examinees and items.

A closer look at small sample sizes

Within sample size N=300, the two concurrent conditions resulted in the smallest linking error in item and ability parameters. There was no significant difference between

20 and 30 common items within the concurrent condition. When these two optimal N=300 conditions were compared to N=500 conditions, on some outcomes the best N=300 conditions showed smaller linking error than the worst N=500 conditions (e.g., theta RMSE). In general, as expected, the best N=500 conditions outperformed the best N=300 conditions, with significant differences in absolute theta bias, absolute difficulty bias, and absolute discrimination bias. Other outcomes showed no significant differences. On the majority of outcomes for sample sizes of 300 or 500, the concurrent condition minimized linking error, without significant differences between 20 and 30 common items.

Dropped replications

The separate calibration condition resulted in many more dropped replications than the concurrent calibration condition ($\chi^2(1)=21.22$, $p<0.001$). Conditions with 300 subjects per group demonstrated the highest drop rate (40/800 replications; Table 2). Surprisingly, the conditions with a higher proportion of linking items showed a higher number of dropped replications for almost every sample size, although this result was not statistically significant ($\chi^2(1)=1.33$, $p=0.25$).

Discussion

This simulation study assessed the feasibility of using small sample sizes and a large proportion of common items to develop a large CAT item bank. A significant improvement in several endpoints was observed for 500 examinees compared to 300, and the improvement was often equal to or greater than the improvement between 500 and

1000 examinees. At sample sizes of N=300 and 500, the concurrent calibration conditions outperformed or showed negligible difference from separate calibration with Stocking Lord linking. For 300 and 500 examinees per group, the one-third common item condition performed only slightly worse than the one-half condition, while increasing the total items calibrated from 270 to 340; in a minimal resource setting, the substantial increase in item bank size may outweigh the relatively small decrease in parameter recovery accuracy.

Sample size

The use of small samples to develop a large CAT item pool could make CAT development more accessible to small, non-commerical research groups. Few studies have investigated the feasibility of N≤500 in a common-items nonequivalent groups design, despite interest in small sample sizes. This study investigated the feasibility of small samples (N=300 per group) under ideal conditions: 50% common items, the same common items administered to all examinees, >50-item test length, low likelihood of construct drift, and concurrent calibration. Even under such favorable conditions, the 300 sample size showed high levels of error. The sample size of 500 examinees, however, shower lower and more acceptable error, particularly in absolute bias of theta, all, items, and common items. Thus, a minimum sample of 500 may be necessary to achieve more tolerable levels of bias and error.

Calibration method

Separate calibration followed by Stocking-Lord linking slightly outperformed concurrent calibration in theta recovery for large sample sizes (N=1000 and 3000 per

group). However, in all other regards, results support the use of concurrent calibration. First and foremost, separate calibration resulted in more difficulties with model fit. For N=300 examinees, nine percent of replications dropped in the separate calibration condition; this high incidence of model failure is undesirable in a low resource setting where data cannot easily be replaced. Of the 35 dropped replications, only one model failed to converge; the remaining 34 replications were dropped because individual items could not be calibrated and would require pruning in an empirical study. Thus, researchers employing separate calibration with a small sample are recommended to include more items to allow for pruning to improve model fit. Separate calibration may show more dropped replications than concurrent calibration because it requires fitting eight times more models (one per group), so there are simply more opportunities for model fit issues to arise.

Concurrent methods showed small but clear advantages over separate calibration in signed theta bias (e.g., 0.03 improvement for N=300). The concurrent conditions slightly underestimated thetas (mean -0.004), while the separate conditions more substantially overestimated thetas (mean 0.020). The concurrent conditions also outperformed the separate conditions in common item parameter bias and theta RMSE.

Nonequivalent groups

The group with the largest difference from the reference group N(0,1) was group 18-19 years N(-0.83, 0.86), which showed significantly higher bias than all other groups. This finding is likely due to the difference in both mean and standard deviation, which made scoring more challenging. Other studies of nonequivalent groups have not reported

comparable bias, even among studies with larger mean differences between groups (e.g., Li & Lissitz, 2012; Kang & Petersen, 2012). Further research could explore the boundaries of this issue of significant bias in nonequivalent groups.

Limitations and future directions

One limitation concerns group distributions, which were empirically derived from an existing IQ assessment to represent change in cognitive ability over adulthood. More systematic variation in mean and standard deviation may have improved the applicability of these results to other projects. However, we believe that the results contained herein are likely to apply to development of other adult scales. Educational achievement assessments routinely employ multigroup methodology, but adult assessments often treat adults as a unitary group with a normally distributed latent trait. Many latent traits besides IQ change between ages 18 and 70, and small between-group differences could be accounted for with multgroup designs such as this one.

In addition, the common item proportions were both fairly moderate (1/2 and 1/3). Previous studies of other study conditions have demonstrated the potential feasibility of fewer common items. A large difference in common items results in a large difference in item bank size. In the present study, 350 verbal items were available, and the number of groups was fixed at 8; accordingly, common item proportions of 20 and 30 were selected, which resulted in total item banks of 270 and 340 items, respectively. Further simulations that alter other variables affecting item bank size (i.e., number of groups and test length) could investigate whether fewer common items are necessary to achieve the desired bank size and tolerable bias level.

Conclusions

Specific recommendations are not appropriate because a future developer's selection of study conditions will depend on their tolerance for error and bias. With regard to the present study, an optimal balance between bias and number of examinees was achieved with a) concurrent calibration, which generally outperformed separate calibration, b) 20 common items, as 30 common items offered minimal improvement in bias but a significant decrease in total item pool, and c) 500 examinees per group. Smaller sample sizes save valuable resources but the relationship between additional subjects and error is nonlinear. Although the 300 sample size resulted in high levels of error, the addition of just 200 examinees per group led to large improvements in many outcomes and brought linking error to tolerable levels. The resulting IQ CAT will strengthen current efforts to understand the genetic basis of IQ through improved reliability and validity.

References

Arai, S., & Mayekawa, S. I. (2011). A comparison of equating methods and linking designs for developing an item pool under item response theory. *Behaviormetrika*, 38(1), 1-16.

Bates, T. C., Lewis, G. J., & Weiss, A. (2013). Childhood socioeconomic status amplifies genetic effects on adult intelligence. *Psychological Science*, *24*(10), 2111-2116.

Betz, N. E. (1977). Effects of immediate knowledge of results and adaptive testing on ability test performance. *Applied Psychological Measurement*, *1*(2), 259–266. doi:10.1177/014662167700100212

Braveman, P. A., Cubbin, C., Egerter, S., Chideya, S., Marchi, K. S., Metzler, M., & Posner, S. (2005). Socioeconomic status in health research: one size does not fit all. *JAMA*, *294*(22), 2879-2888.

Calvin, C. M., Deary, I. J., Webbink, D., Smith, P., Fernandes, C., Lee, S. H., ... & Visscher, P. M. (2012). Multivariate genetic analyses of cognition and academic achievement from two population samples of 174,000 and 166,000 school children. *Behavior Genetics*, *42*(5), 699-710.

Davies, G., Marioni, R. E., Liewald, D. C., Hill, W. D., Hagenaars, S. P., Harris, S. E., ... & Cullen, B. (2016). Genome-wide association study of cognitive functions and educational attainment in UK Biobank (N= 112 151). *Molecular Psychiatry*, *21*(6), 758.

Deary, I. J., Corley, J., Gow, A. J., Harris, S. E., Houlihan, L. M., Marioni, R. E., ... & Starr, J. M. (2009). Age-associated cognitive decline. *British Medical Bulletin*, *92*(1), 135-152.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, 22, 144-149.

Han, K. T. (2007). WinGen: Windows software that generates item response theory parameters and item responses. *Applied Psychological Measurement*, 31(5), 457-459.

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, *26*(1), 3-24.

Hedden, T., & Gabrieli, J. D. (2004). Insights into the ageing mind: a view from cognitive neuroscience. *Nature reviews neuroscience*, *5*(2), 87.

Intelligence Research Should Not Be Held Back by its Past. (2017). *Nature.* 545, 385–386. doi:10.1038/nature.2017.22021

Kang, T., & Petersen, N. S. (2012). Linking item parameters to a base scale. *Asia Pacific Education Review*, *13*(2), 311-321.

Kim, S. H., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*, *22*(2), 131-143.

Kim, S., & Kolen, M. J. (2004). STUIRT [Computer software]. Iowa City, IA: The Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available from the web address: http://www.uiowa.edu/~casma)

Kim, S. & Kolen, M. J. (2006) Robustness to Format Effects of IRT Linking Methods for Mixed-Format Tests, *Applied Measurement in Education, 19*(4), 357-381, doi: 10.1207/s15324818ame1904_7

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: methods and practices* (3rd ed.). New York: Springer.

Knopik, V. S., Neiderhiser, J. M., DeFries, J. C., & Plomin, R. (2016). *Behavioral Genetics* (7th ed.). New York: Worth Publishers.

Krasileva, K. E., Sanders, S. J., & Bal, V. H. (2017). Peabody Picture Vocabulary Test: Proxy for verbal IQ in genetic studies of autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *47*(4), 1073-1085.

Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., ... & Fontana, M. A. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics*, *50*(8), 1112.

Li, Y., & Lissitz, R. W. (2012). Exploring the full-information bifactor model in vertical scaling with construct shift. *Applied Psychological Measurement*, *36*(1), 3-20.

Lord, F. M. (1968). Some test theory for tailored testing. *ETS Research Bulletin Series*, *1968*(2), i-62.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch Model. *Journal of Education Measurement*, *17*, 179-193.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, *14*, 139-160.

Plomin, R., & von Stumm, S. (2018). The new genetics of intelligence. *Nature Reviews Genetics*, *19*(3), 148.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rietveld, C. A., Esko, T., Davies, G., Pers, T. H., Turley, P., Benyamin, B., ... & De Leeuw, C. (2014). Common genetic variants associated with cognitive performance identified using the proxy-phenotype method. *Proceedings of the National Academy of Sciences*, *111*(38), 13790-13794.

Savage, J. E., Jansen, P. R., Stringer, S., Watanabe, K., Bryois, J., De Leeuw, C. A., ... & Grasby, K. L. (2018). Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nature Genetics*, *50*(7), 912.

Spearman, C. (1904). 'General Intelligence,' objectively determined and measured. *The American Journal of Psychology, 15*(2), 201-292.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201-210.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. Mahwah, NJ: Lawrence Earlbaum Associates.

Wang, S., McCall, M., Jiao, H., & Harris, G. (2012). Construct Validity and Measurement Invariance of Computerized Adaptive Testing: Application to

Measures of Academic Progress (MAP) Using Confirmatory Factor Analysis. Paper presented at *The Annual Meeting of the American Educational Research Association (AERA)*. Vancouver, British Columbia: Canada.

Wechsler, D. (2011*). Wechsler Abbreviated Scale of Intelligence–Second Edition (WASI-II)*. San Antonio, TX: NCS Pearson.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV): Technical and Interpretative Manual*. San Antonio, TX: Pearson.

Weiss, D. J. (2004). Computerized Adaptive Testing for Effective and Efficient Measurement in Counseling and Education. *Measurement and Evaluation in Counseling and Development*, *37*(2), 70–84.

Zabaneh, D., Krapohl, E., Gaspar, H. A., Curtis, C., Lee, S. H., Patel, H., ... & Lubinski, D. (2018). A genome-wide association study for extremely high intelligence. *Molecular Psychiatry*, *23*(5), 1226.

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). BILOG-MG (version 3.0): Multiple-group IRT analysis and test maintenance for binary items [Computer software manual]. Skokie, IL: Scientific Software International, Inc.

Tables

Table 1.

Nonequivalent groups

Group	Age group	Mean	SD
1	18-19	-0.3	0.86
2	20-24	-0.2	0.89
3	25-29	-0.1	0.93
4	30-34	0.0	1.00
5	35-44	0.0	1.04
6	45-54	0.0	1.07
7	55-64	0.0	1.14
8	65-70	0.1	1.18

Table 2.

Theta recovery

								By group θ signed bias							
N Subjects	n Com-mon Items	Calibra-tion	n Dropped replica-tions	θ mean SE	θ mean RMSE	θ absolute bias	θ signed bias	18-19	20-24	25-29	30-34	35-44	45-54	55-64	65-70
300	20	C	2	1.047	1.079	0.254	-0.001	-0.179	0.023	0.022	0.021	0.024	0.023	0.023	0.030
		S	15	1.072	1.102	0.253	0.029	0.011	0.029	0.027	n/a	0.034	0.034	0.030	0.037
	30	C	3	1.049	1.081	0.252	-0.003	-0.186	0.018	0.018	0.024	0.025	0.022	0.028	0.031
		S	20	1.071	1.102	0.251	0.023	-0.011	0.025	0.023	n/a	0.031	0.030	0.032	0.029
500	20	C	0	1.057	1.088	0.251	-0.006	-0.184	0.019	0.017	0.017	0.019	0.021	0.020	0.028
		S	2	1.076	1.105	0.250	0.021	0.008	0.021	0.019	n/a	0.025	0.026	0.021	0.030
	30	C	1	1.052	1.082	0.247	-0.003	-0.187	0.018	0.019	0.023	0.022	0.023	0.025	0.030
		S	3	1.065	1.094	0.245	0.019	-0.017	0.022	0.023	n/a	0.025	0.025	0.026	0.025

1000	20	C	0	1.055	1.085	0.246	-0.006	-0.184	0.018	0.019	0.017	0.017	0.021	0.019	0.024
		S	0	1.067	1.095	0.244	0.020	0.015	0.019	0.019	n/a	0.018	0.023	0.019	0.025
	30	C	0	1.052	1.082	0.245	-0.004	-0.186	0.018	0.019	0.020	0.021	0.022	0.025	0.027
		S	1	1.057	1.084	0.240	0.015	-0.031	0.020	0.021	n/a	0.020	0.021	0.023	0.022
3000	20	C	1	1.060	1.089	0.243	-0.007	-0.184	0.016	0.017	0.016	0.017	0.019	0.021	0.024
		S	0	1.069	1.097	0.241	0.021	0.046	0.017	0.016	n/a	0.016	0.017	0.019	0.021
	30	C	1	1.056	1.085	0.243	-0.005	-0.190	0.016	0.018	0.020	0.022	0.022	0.024	0.027
		S	0	1.055	1.082	0.237	0.014	-0.023	0.018	0.019	n/a	0.020	0.019	0.019	0.020

θ absolute bias

300	20	C						0.321	0.232	0.236	0.235	0.240	0.250	0.255	0.261
		S						0.296	0.239	0.241	n/a	0.243	0.252	0.256	0.262
	30	C						0.322	0.225	0.231	0.236	0.240	0.248	0.252	0.261
		S						0.296	0.232	0.237	n/a	0.242	0.252	0.255	0.259
500	20	C						0.321	0.227	0.232	0.233	0.237	0.245	0.252	0.258
		S						0.298	0.232	0.237	n/a	0.240	0.247	0.252	0.258

	30	C	0.320	0.222	0.227	0.232	0.234	0.243	0.246	0.253
		S	0.290	0.227	0.231	n/a	0.236	0.244	0.247	0.251
1000	20	C	0.319	0.222	0.228	0.229	0.232	0.240	0.246	0.251
		S	0.297	0.224	0.231	n/a	0.233	0.239	0.245	0.248
	30	C	0.320	0.219	0.225	0.230	0.232	0.239	0.243	0.249
		S	0.283	0.222	0.227	n/a	0.231	0.239	0.242	0.245
3000	20	C	0.321	0.220	0.225	0.227	0.230	0.237	0.242	0.246
		S	0.312	0.221	0.226	n/a	0.228	0.235	0.239	0.242
	30	C	0.321	0.218	0.223	0.228	0.230	0.238	0.240	0.245
		S	0.286	0.219	0.224	n/a	0.227	0.235	0.237	0.240

Table 3.

Parameter recovery

			Absolute bias			Signed bias			Absolute bias, common items		
N Subjects	n Common Items	Calibration	*a*	*b*	*c*	*a*	*b*	*c*	*a*	*b*	*c*
300	20	C	0.203	0.184	0.027	0.014	-0.005	0.002	0.123	0.121	0.027
		S	0.208	0.185	0.028	-0.049	-0.027	-0.013	0.193	0.191	0.025
	30	C	0.198	0.178	0.026	0.003	-0.010	-0.001	0.112	0.112	0.027
		S	0.204	0.183	0.027	-0.049	-0.035	-0.013	0.184	0.175	0.027
500	20	C	0.180	0.154	0.027	-0.010	-0.007	0.003	0.105	0.099	0.025
		S	0.187	0.156	0.028	-0.056	-0.025	-0.011	0.173	0.160	0.025
	30	C	0.175	0.145	0.027	-0.008	-0.002	0.001	0.091	0.091	0.027
		S	0.180	0.149	0.028	-0.045	-0.023	-0.011	0.163	0.146	0.027

1000	20	C	0.151	0.120	0.027	-0.022	-0.002	0.004	0.080	0.078	0.023
		S	0.156	0.118	0.028	-0.051	-0.012	-0.008	0.137	0.117	0.024
	30	C	0.148	0.117	0.027	-0.021	0.000	0.001	0.073	0.076	0.024
		S	0.150	0.115	0.028	-0.039	-0.013	-0.009	0.127	0.107	0.026
3000	20	C	0.118	0.086	0.025	-0.044	-0.008	0.002	0.067	0.056	0.020
		S	0.120	0.083	0.026	-0.056	-0.004	-0.005	0.092	0.074	0.022
	30	C	0.116	0.087	0.025	-0.038	-0.004	-0.001	0.061	0.058	0.020
		S	0.112	0.082	0.026	-0.040	-0.004	-0.006	0.085	0.073	0.025

Figures

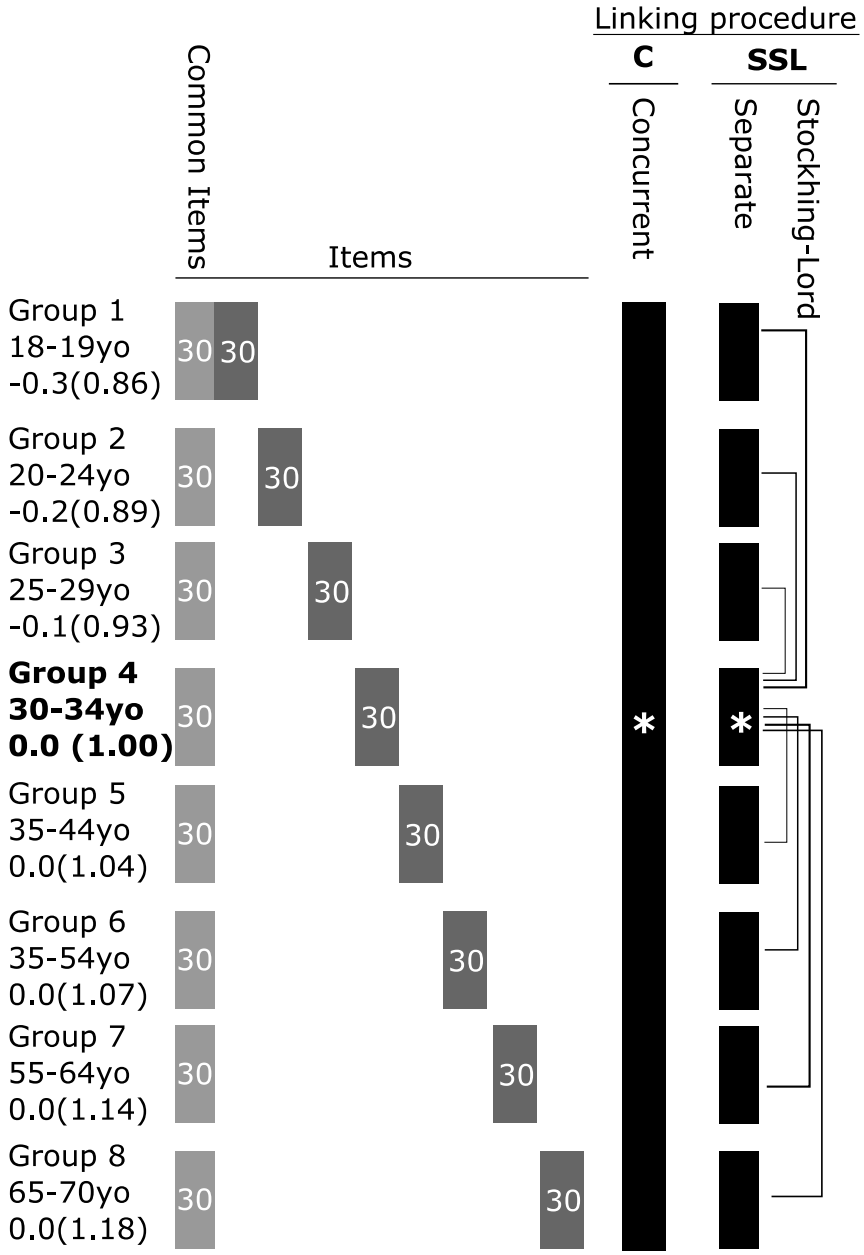

Figure 1. Nonequivalent groups Anchor Test (NEAT) design for 30 common item conditions. Group 4 (in bold and with asterisks) served as the reference group. The approximated age range (e.g., 30-34 year olds) and distribution (e.g., mean=0.0 and standard deviation=1.00) were derived from the Wechsler scoring tables.

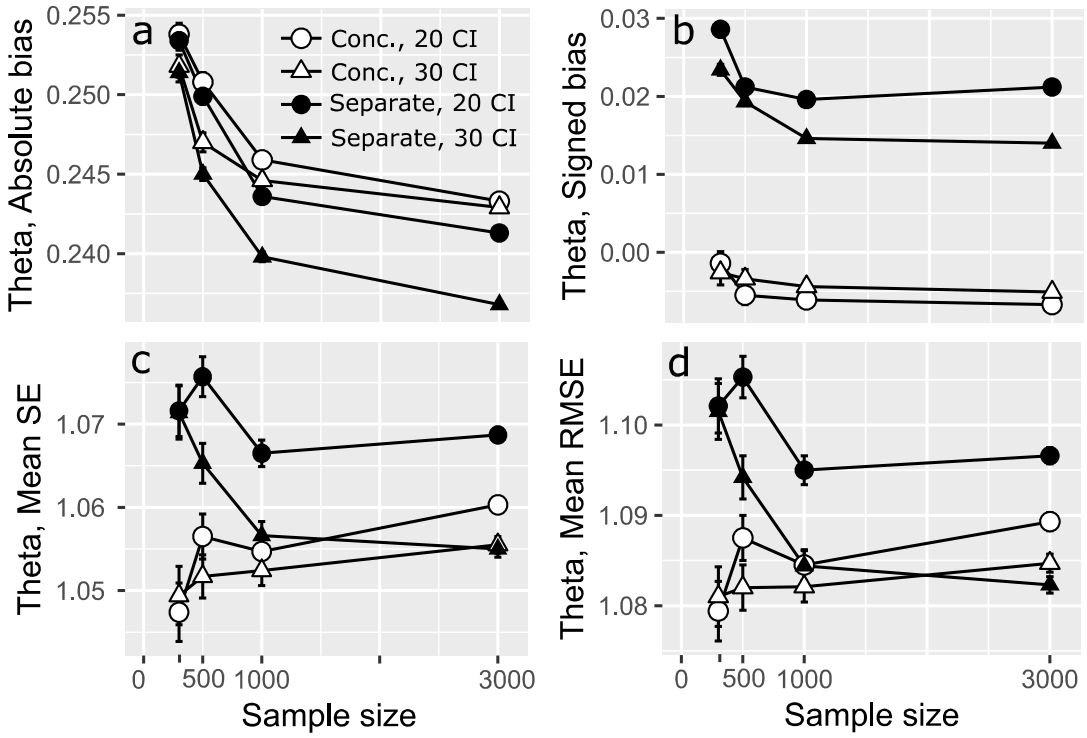

Figure 2. Theta bias and error

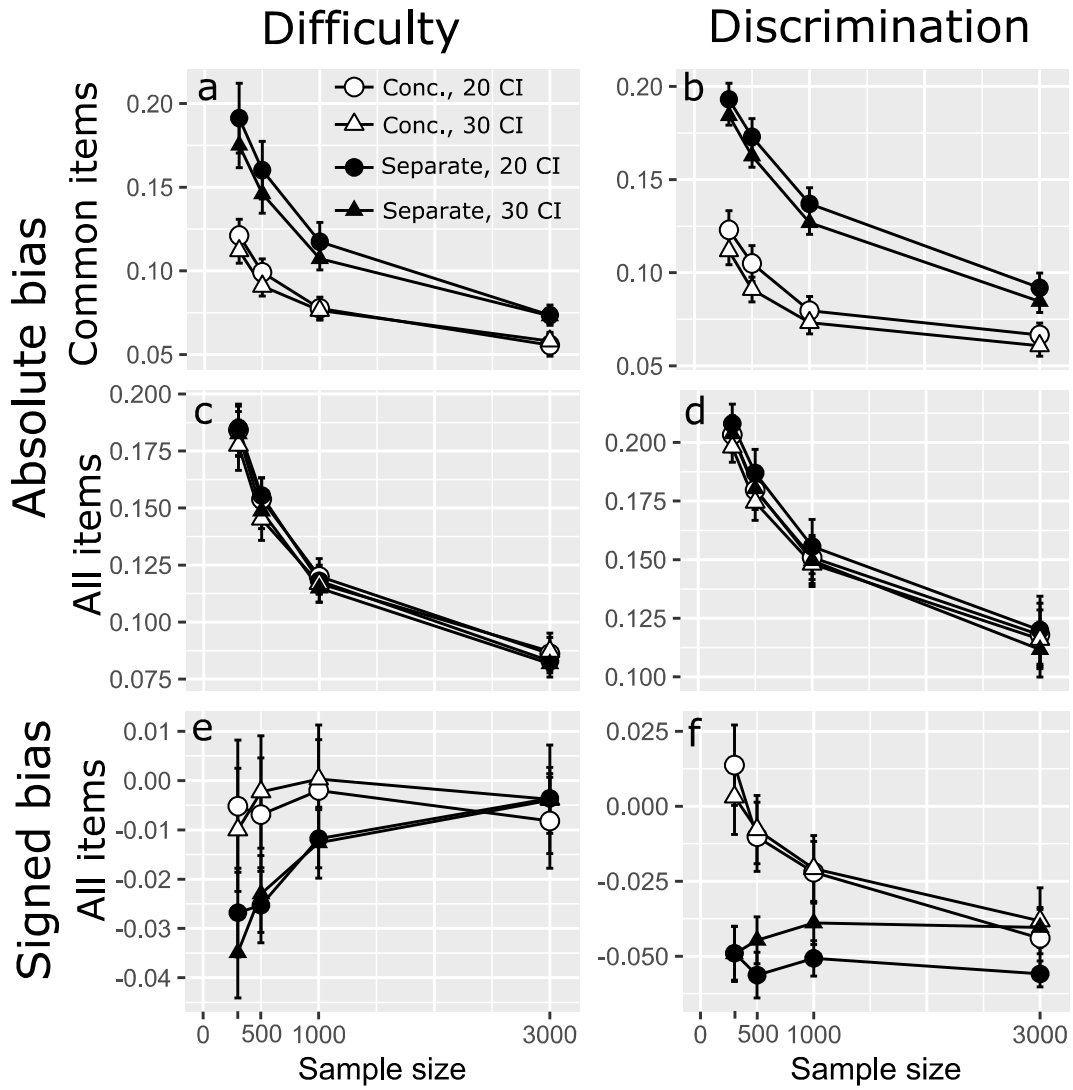

Figure 3. Item parameter bias

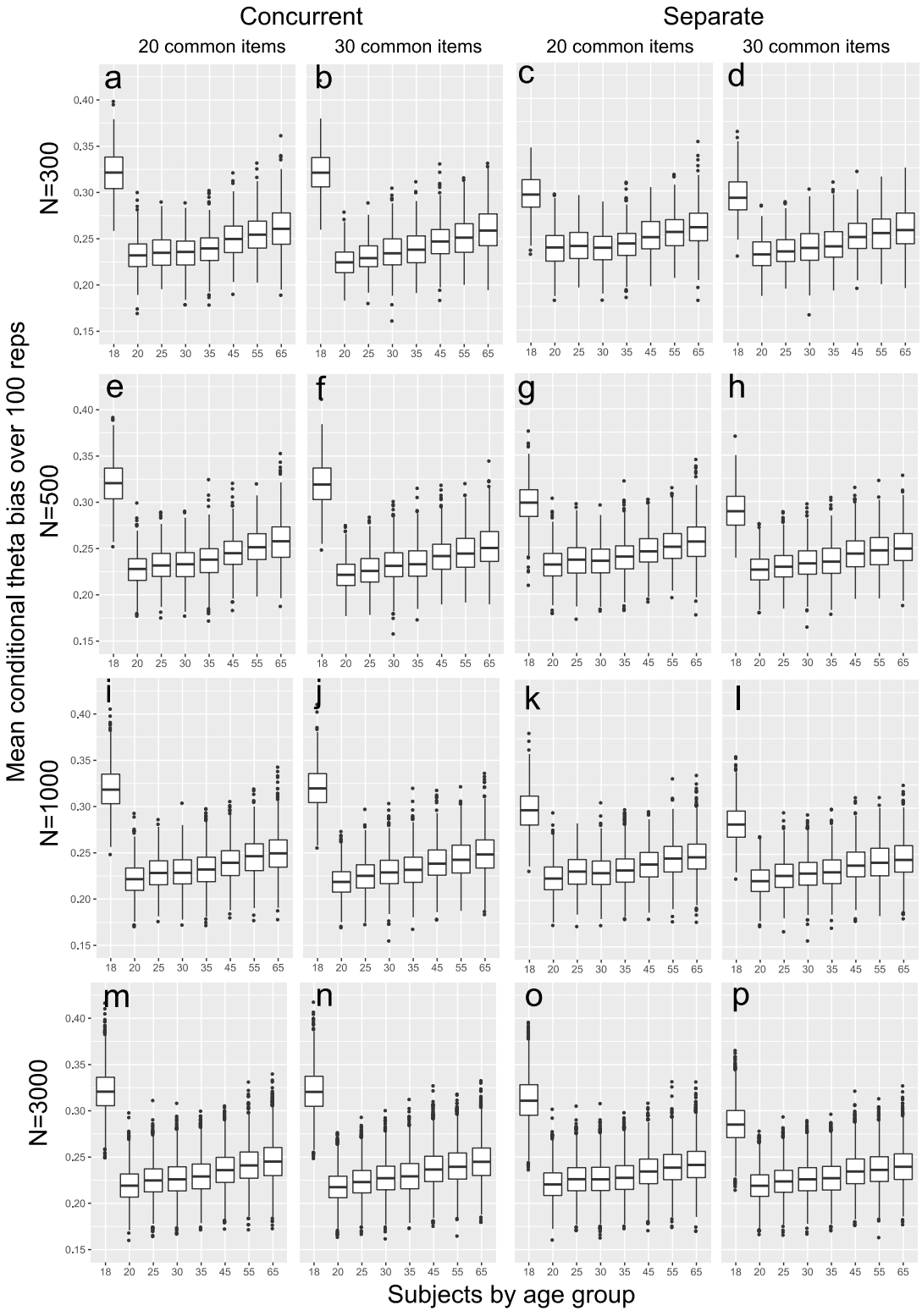

Figure 4. Absolute theta bias by group

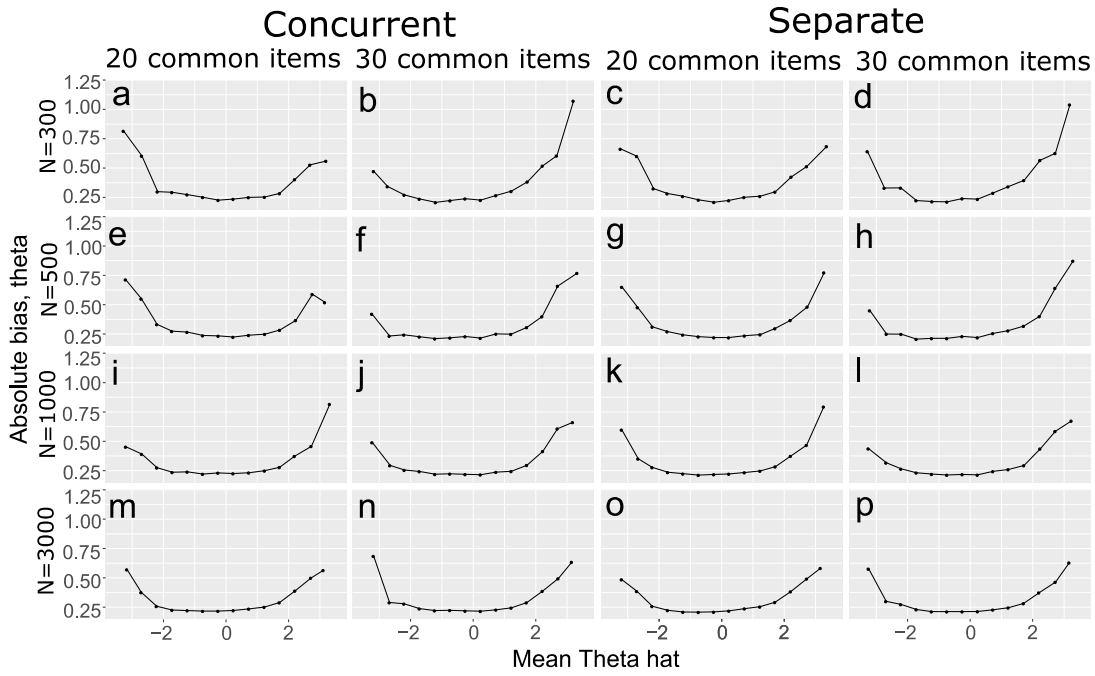

Figure 5. Mean absolute theta bias by binned ability level, replication 001

CHAPTER 3: DOES THE FACTOR STRUCTURE OF IQ DIFFER BETWEEN THE DAS-II NORMATIVE SAMPLE AND AUTISTIC CHILDREN?

This work originally appeared in Autism Research (2020).

Key Words: autism spectrum disorders, intelligence, educational psychology, factor analysis, validity

Abstract

The Differential Abilities Scales, 2nd edition (DAS-II) is frequently used to assess intelligence in autism spectrum disorder (ASD). However, it remains unknown whether the DAS-II measurement model (e.g., factor structure, loadings), which was developed on a normative sample, holds for the autistic population or requires alternative score interpretations. We obtained DAS-II data from 1,316 autistic individuals in the Simons Simplex Consortium and 2,400 individuals in the normative data set. We combined ASD and normative data sets for multigroup confirmatory factor analyses to assess different levels of measurement invariance, or how well the same measurement model fit both data sets: "weak" or metric, "strong" or scalar, and partial scalar if full scalar was not achieved. A weak invariance model showed excellent fit (Confirmatory Fit Index [CFI] > 0.995, Tucker Lewis Index [TLI] > 0.995, root mean square error of approximation [RMSEA] < 0.025), but a strong invariance model demonstrated a significant deterioration in fit during permutation testing (all p's<0.001), suggesting measurement bias, meaning systematic error when assessing autistic children. Fit improved significantly, and partial scalar invariance was achieved when either of the two spatial subtest (Recall of Designs or Pattern Construction) intercepts was permitted to vary between the ASD and normative groups, pinpointing these subtests as the source of bias. The DAS-II appears to measure verbal and nonverbal-but not spatial-intelligence in autistic children similarly as in normative sample children. These results may be driven by Pattern Construction, which shows higher scores than other subtests in the ASD sample. Clinicians assessing autistic children with the DAS-II should interpret verbal and

nonverbal reasoning composite scores over the spatial score or General Composite Ability.

Abbreviations

ADHD: Attention Deficit/Hyperactivity Disorder; AIC: Akaike information criterion; ASD: Autism Spectrum Disorder; CFI: Confirmatory Fit Index; DAS-II: Differential Abilities Scales, 2nd Edition; ID: Intellectual Disability; IQ: Intelligence quotient; GCA: General Conceptual Ability; NVIQ: Nonverbal IQ; RMSEA: Root Mean Square Error of Approximation; SD: Standard deviation; Seq. & quant Reasoning: Sequential & quantitative reasoning; SNC: Special Nonverbal Composite; SRMR: Standardized Root Mean Square Residual; SSC: Simons Simplex Consortium; TLI: Tucker Lewis Index; VIQ: Verbal IQ

Introduction

Intellectual disability (ID) commonly co-occurs with autism spectrum disorder (ASD): approximately 50% of autistic individuals meet criteria for ID (Charman et al., 2011). To assess ID in school-age autistic children, clinicians frequently use the DAS-II (Differential Ability Scales, 2nd Edition, Elliott, 2007a) to measure cognitive ability. However, it remains unknown whether the DAS-II functions similarly in autistic and neurotypical children (Wicherts, 2016). The DAS-II measurement model (i.e., the relationship between subtests and the latent constructs of verbal, nonverbal, and spatial intelligence which is described by the factor structure, factor loadings, covariances, etc.) was developed with a nationally representative normative sample, and has never been tested in a large autistic sample to our knowledge. If the DAS-II measurement model fails to hold for autistic children, alternative methods and score interpretations will be needed for measuring cognitive ability and informing ID assessments.

Research has shown that the measurement models of some intellectual assessments perform differently in some subgroups. For example, the DAS-II measurement model showed small differences for a sample of African Americans (Trundt et al., 2018), the WISC-IV measurement model showed differences for a sample with ADHD (Thaler et al., 2015), and a factor analysis of the WAIS-R, WAIS-III, WISC-R, and WISC-III in a sample of high functioning autism identified a 'social context' factor not present in the normative sample (Goldstein et al., 2008). When a measurement model performs differently in a particular subgroup, this suggests that measurement bias affects scores for individuals in that subgroup such that their measured scores do not reflect their

true scores on the latent trait (e.g., nonverbal intelligence) in the same way that scores for the normative group do, whether driving measured scores up or down (Reynolds & Lowe, 2009). Please note that throughout this article, the terms 'nonverbal intelligence' or NVIQ are used instead of fluid reasoning (g_f) for consistency with DAS-II nomenclature (Elliott et al., 2018, p. 347).

Clinicians have long discussed "IQ splits" in individuals with ASD, and recent research lends more support to this observed phenomenon. Siegel and colleagues (1996) initially reported that in 45 high-functioning autistic individuals, 36% of participants showed unusually large differences (i.e., 12 IQ points in standard scale of mean 100, *SD* 15) between their nonverbal IQ and verbal IQ scores (20% NVIQ > VIQ, 16% VIQ > NVIQ). Many other researchers reported similar data, and an analysis of the largest known sample of DAS-II data on autistic children (n = 2,110; the Simons Simplex Consortium) confirmed the 'splits' finding with 32% of individuals showing DAS-II Early Years NVIQ > VIQ discrepancies of at least 16 points, and 20% showing the same discrepancy on DAS-II School Age (Nowell et al., 2015). At present, it is unclear whether these 'splits' reflect true differences between verbal and nonverbal intelligence, or are better attributed to measurement bias due to a poor fit of the DAS-II measurement model in autistic children. This question can be answered by testing measurement invariance.

Measurement invariance is a method to determine whether an assessment such as the DAS-II measures the same latent construct with the same precision in multiple populations. In other words, it tests whether the observed test score of an individual -

who has a certain true score on the latent construct - is independent of that individual's group membership (Thompson, 2016). Different levels of measurement invariance are tested sequentially with increasing strictness. At the first level, the *same confirmatory factor model* is fit to each group separately. This level of invariance merely demonstrates that the same model can be fit to each group, but does not rule out measurement bias in the relationship between one group's test scores and true ability. At the second or "weak" factorial invariance level, configural invariance, a multigroup model is fit to the combined datasets; this model requires that the *same items load on the same factors* for each group, but imposes no between-group constraints on factor loadings or any other parameters. At the third level, also referred to as "weak" factorial invariance, *factor loadings are constrained* to be equal in both groups, but no other between-group constraints are imposed. At the fourth level, scalar or "strong" factorial variance is required to conclude that between-group differences in mean scores are entirely due to true group differences in latent abilities and not measurement bias. Scalar invariance requires *equality between groups on intercepts*, and permits estimation of differences between group factor means by no longer setting factor means equal to 0 as in metric and configural invariance. In one final level, residual or "strict" invariance, *residuals are constrained* to be equal in both groups. However, this level of factorial invariance is not necessary; it is widely accepted that scalar or "strong" invariance is sufficient for use of a measure with a particular population, such as autistic children. If scalar invariance is achieved between the autistic and normative samples, then it can be concluded that group differences in nonverbal, verbal, and spatial intelligence scores reflect true group

differences in ability. If scalar invariance is not achieved, then group differences might be due to measurement bias and artifacts rather than true differences in intelligence; thus an autistic child's DAS-II score would be biased compared to the normative sample.

The objective of this study is to determine whether DAS-II scores are biased for autistic children.

Methods

Participants

The ASD sample was drawn from the Simons Simplex Collection (SSC), which was a multi-site study of 2,110 children ages 4-18 who met gold-standard diagnostic criteria for ASD. Participants completed a comprehensive diagnostic and behavioral testing battery that included the DAS-II School Age core subtests. For additional information on SSC data collection, recruitment, diagnoses and inclusion criteria, see Fischbach and Lord, 2010. SSC participants were included in the present study if they had a DAS-II School Years subtest score (n = 1,316; see Table 1). Over 90% of participants had complete data on all six core DAS-II subtests.

The control sample consisted of the nationally representative DAS-II School Age normative sample ages 6-17 (n = 2,400; see Table 1) and was provided by Pearson, publisher of the DAS-II. For additional information on this sample, see the DAS-II Technical Manual (Elliott, 2007b).

This study was approved by The Children's Hospital of Philadelphia Institutional Review Board and adheres to the legal requirements of the United States.

Data Analysis

Missing data. Eight of 2,400 individuals in the normative dataset were missing data on one subtest. The ASD sample showed significantly more missing data (119 of 1,316 participants). While each nonverbal and spatial subtest had data from > 99% of ASD participants, both verbal subtests were missing for 8.1% of participants (n = 106). Data were not missing at random: the 106 participants with verbal subtest missingness showed substantially lower verbal abilities on other measures (Verbal Communication score on the Autism Diagnostic Interview – Revised, t(120) = -7.08, $p < 0.001$, mean missing = 19.0, mean nonmissing = 16.3) and module selected for the Autism Diagnostic Observation Schedule, which is based on language level and age ($\chi^2(3) = 586.0$, $p < 0.001$). The ASD sample showed a very wide ability range with and without these 106 participants, and in fact the range of General Composite Ability remained the same (40-167).

All analyses were conducted on the full datasets that included all participants, including those missing subtest score(s) which were imputed by Full Information Maximum Likelihood, following the guidelines provided by Newman (2014). Then, in an effort to explore any bias introduced by the missing data from 119 ASD participants, we conducted sensitivity analyses to determine whether meaningful differences resulted. First, we reconducted analyses excluding participants with missing data (i.e., listwise deletion). Second, we adjusted imputed values by subtracting and adding arbitrary values (implemented with the mice package in R (van Buuren & Groothuis-Oudshoorn, 2011)), then we reconducted analyses with the new datasets. Third, we tested the base oblique model in the ASD dataset alone using an auxiliary variable related to verbal

communication: the parent report ADI-R (Autism Diagnostic Interview – Revised) verbal communication total score. Auxiliary variable analysis and Full Information Maximum Likelihood were implemented in Mplus v8.2 (Muthen & Muthen, 1998).

Confirmatory factor analysis. First, we determined the base model for invariance testing by fitting the same confirmatory factor model separately to the normative data and to the ASD data to ensure the most basic measurement model fit both samples. In selecting the target model, we consulted the DAS-II Technical Manual, which reported two models. The first model, a correlated three-factor ("oblique") model, uses the 6 core subtests, similar to our dataset (Elliott, 2007b, p. 159). The correlated three-factor model allows correlations between the 3 factors (verbal, nonverbal reasoning, and spatial) and does not include a higher order general (*g)* factor (Figure 1). The Technical Manual describes a second model, the higher-order model, which uses both the 6 core subtests and the less frequently used 6 diagnostic subtests (Elliott, 2007b, p. 157). In the higher-order model, the 6 core subtests load onto 3 factors (verbal, nonverbal reasoning, and spatial), which in turn load onto a general (*g*) factor; the diagnostic subtests load onto 3 separate factors that in turn load onto *g* (Figure S1). Of note, for the 6-core subtest battery, the Technical Manual reports fit statistics for the correlated three-factor and not the higher-order model. The Technical Manual does not describe fitting the higher-order model to the 6-core subtest battery alone, which is most commonly used clinically and in our ASD dataset. Note that we use the classical definition of the term 'higher-order model' to refer to the model in Figure S1, which is sometimes called by the name of the more general category to which it belongs, 'hierarchical model.'

In addition, we fit bifactor models demonstrated by previous research to fit the normative data (e.g., Canivez & McGill, 2016; Dombrowski, Golay, McGill, & Canivez, 2018; Dombrowski, McGill, Canivez, & Peterson, 2019). A bifactor model includes the general factor and group factors (i.e., verbal, nonverbal, and spatial) and assumes that the general factor is orthogonal to the group factors. Note that we use the classical definition of the term 'group factor' to refer to verbal, nonverbal, and spatial factors (sometimes referred to as specific factors). We fit two bifactor models: a 3-factor bifactor model with verbal, nonverbal, and spatial group factors and *g* as suggested by Canivez and McGill (2016), and a 2-factor bifactor model with verbal and spatial group factors, and the two nonverbal reasoning subtests loading directly on *g* instead of a nonverbal factor (Figure S2) as reported by Dombrowski et al. (2018). In the bifactor models, we fixed correlations between all factors at 0, and fixed equality between the two factor loadings on each group factor, which decreases the number of parameters being estimated and thus allows model identification. Finally, we also fit a simple unidimensional model that allowed the six subtests to load directly on *g*.

Measurement invariance. Next, we combined the normative and ASD datasets into one multigroup dataset. We used the best model established in the previous step to test sequentially stricter levels of measurement invariance: configural, metric (weak), scalar (strong), then residual (strict). If invariance was not achieved, we ran partial invariance tests to identify the locus of misfit.

Comparisons between measurement invariance models were made in accordance with recommendations by Jorgensen and colleagues (2018) to assess statistical

significance rigorously via permutation testing, rather than cut-offs established by Chen (2007), which have inconsistent Type I error rates. In each permutation, group membership was randomly assigned; a distribution was built from 1000 replications then used to determine whether true group membership differed significantly from what would be expected under the null hypothesis, as evidenced by the size of change among fit indices during the replications. We rejected models with $p < 0.05$ on multiple fit indices in favor of the simpler model in the comparison.

All primary factor analyses were implemented in Mplus version 8.2 (Muthen & Muthen, 1998). Permutation testing, effect size estimation, sensitivity analyses, and all remaining analyses were implemented in R version 3.5.2 (R Core Team, 2018) using packages lavaan (Rosseel, 2012) and semTools (Jorgensen, Pornprasertmanit, Schoemann, & Rosseel, 2019). All models were estimated with maximum likelihood with robust standard errors (implemented with MLR) due to significant non-normality of every subtest in both datasets according to the Shapiro-Wilk test (all $W > 0.95$, all $p < 0.02$).

Results

Confirmatory Factor Analysis

First we fit a correlated 3-factor model as reported in the Technical Manual for the 6-subtest core battery. The model demonstrated excellent fit with the data, as expected. See tables 2, 3, and S1 for complete fit statistics for all eight models and intersubtest correlations. The higher-order model yielded a factor loading > 1.0 of the nonverbal factor on *g* for both the normative (1.005) and ASD (1.045) datasets (Table

S1). These results suggest that the nonverbal factor contributes no specific variance. In other words, the general factor absorbs all variance in the nonverbal factor. We next attempted to fit a three-factor bifactor model that allows subtests to load on both group and general factors. The three-factor bifactor model did not converge for either the normative or ASD datasets. After removing individuals with missing data, however, the model converged for the normative dataset and yielded a factor loading of 1.00 for the nonverbal factor on *g*, indicating persistence of nonverbal factor variance issues. Additionally, the three-factor bifactor model did not converge at all for the ASD dataset. A bifactor model with two group factors (verbal and spatial) and *g* loaded by all six core subtests (i.e., the nonverbal subtests did not form a general factor) converged for both datasets and showed excellent fit. The unidimensional model demonstrated poor fit for each group (TFI and CLI < 0.94 for both ASD and Norm groups; Table 3) and did not merit further exploration.

An acceptable base solution must be adequate in terms of both model fit and psychological interpretation (Jöreskog, 1969). The higher-order and bifactor models with three first-order and group factors, respectively, were psychologically interpretable but produced improper solutions or failed to converge (Diamantopoulos & Siguaw, 2003), making them inappropriate for consideration as base models for invariance testing. The three-factor oblique model and the two-factor bifactor model without a separate nonverbal group factor both exhibited acceptable model fit although the oblique model showed slightly better fit than the two-factor bifactor model on many indices (χ^2, *p*, CFI, TLI, SRMR, and for ASD only, RMSEA), particularly for the ASD dataset. Although

the two-factor bifactor model could reasonably be selected as the base model, the correlated three-factor model was identified by the publisher, and its results are more easily interpreted by clinicians because the three factors translate directly to the three DAS-II composite scores of verbal, nonverbal reasoning, and spatial, whereas the two-factor bifactor model lacks the nonverbal reasoning factor. Consequently, the correlated three-factor model was chosen as the base model to test measurement invariance between the ASD and normative groups. The final model was tested with the ADI-R verbal communication auxiliary variable, and results did not change meaningfully.

Measurement Invariance

Full invariance. We used the correlated three-factor model to test measurement invariance between the ASD and normative groups. Results indicated that configural and metric invariance were achieved (see table 3 and 5). Scalar invariance was not achieved: on all fit indices, permutation testing showed a significant deterioration in fit (all *p*'s < 0.001, Table 5). Traditional metrics also provided evidence of poor scalar model fit: CFI, TLI, and RMSEA showed change beyond acceptable limits and RMSEA rose over the 0.05 threshold (Chen, 2007).

Partial invariance. Partial scalar invariance was assessed by allowing single subtest intercepts to vary between groups. We observed little change in the model when verbal factor subtests (word definitions or verbal similarities) or nonverbal factor subtests (matrices or sequential and quantitative reasoning) were allowed to vary, suggesting that the model easily accommodates equality between ASD and normative intercepts on these subtests; group differences in verbal and nonverbal factor scores are due to true group

differences in verbal and nonverbal abilities, not bias. This pattern was not true for the subtests loading on the spatial factor (pattern construction and recall of designs). When either of these intercepts was freed to vary between groups, the model fit improved significantly on all indices. This partial scalar invariance model (i.e., with the recall of designs intercept freed) was then tested for partial strict invariance, or holding residuals equal between groups. Partial strict invariance was not achieved (5 of 6 fit indices with *p's*<0.01, Table 5).

A closer look at the full scalar model revealed that the Recall of Designs subtest intercept was 49.3 when held equal between groups; when freed to vary between groups, the Recall of Design intercept was 50.0 for the normative group and 46.6 for the ASD group (Table 4, Figure 2), suggesting that autistic children are expected to have a lower Recall of Designs score than neurotypical children with the same true spatial ability. The opposite pattern was observed for the other spatial subtest, Pattern Construction: when freed, the intercept was 50.0 for the normative group and *increased* to 53.8 for autistic children, indicating that they have a *higher* Pattern Construction score than neurotypical children of the same ability. For comparison, the four verbal and nonverbal subtest intercepts showed much smaller changes, and remained within 0.6 points of the normative group intercept when freed (Table S3). Unlike the verbal and nonverbal factors, spatial factor group differences are not only due to true group differences in spatial ability; some of the difference is also due to measurement bias. For additional data and factor loadings, see supplemental tables.

Factor mean differences. As expected, we observed mean between-group differences on all three factors (Table 4). Autistic children showed unstandardized factor scores that were 0.64, 0.50, and 0.34 lower than normative verbal, nonverbal, and spatial scores, respectively. Unfortunately, we can interpret only the direction, not the size, of these mean differences because they were obtained with the scalar model, which showed a poor fit with the data. The mean factor differences changed in the partial scalar models, but the direction always remained the same.

Missing data. Sensitivity analyses conducted with adjustments to imputed values showed no meaningful differences from primary measurement invariance analyses (i.e., minimal or zero change in fit indices, factor loadings, means, or intercepts).

Discussion

Our findings indicate that the DAS-II School Age measures verbal and nonverbal intelligence in autistic children similarly to how it measures these constructs in neurotypical children, but the same is not true of spatial intelligence. Weak measurement invariance (metric and configural) was achieved for the DAS-II in a multigroup confirmatory factor analysis using a correlated three-factor model, but strong (scalar) measurement invariance was not achieved. Without scalar invariance, group mean differences in DAS-II scores do not reflect true group differences in intelligence alone, but also unique aspects due to being autistic (i.e., measurement bias). Since partial scalar invariance was achieved only when the spatial subtest intercepts were free to vary, we attribute failed scalar invariance to group bias or artifacts in the spatial subtests.

Interpreting DAS-II spatial subtest scores for children with ASD

The two spatial subtests showed large changes in intercepts, in opposite directions, when the intercepts were free to vary between groups. The Recall of Designs intercept for autistic children fell 3.4 points below the normative intercept, while the Pattern Construction intercept rose 3.8 points above the normative intercept. These results indicate that for each subtest, an autistic child's score is expected to be below or above, respectively, the score of a neurotypical child with the same true spatial ability. Simply put, Recall of Designs underestimates an autistic child's true ability, and Pattern Construction overestimates it. The large differences in opposite directions for the spatial subtests should not be interpreted as 'cancelling each other out' because it is likely that different (and unknown) proportions of each subtest's change are due to measurement bias. Although some methods exist for quantifying bias (Nye & Drasgow, 2011), they are more readily applied to unidimensional models than to our three factor model.

The Pattern Construction subtest may be driving the problematic fit: the average autistic participant performed much better on this subtest than on any other. On average, the ASD sample scored around 46 points on all other subtests (45.5-46.7 points; Table 1), but almost 3 points higher on Pattern Construction (48.9 points). In contrast, the normative sample showed nearly identical mean scores on all subtests (50.0-50.2 points). Put another way, the normative sample showed a 0 point difference between Pattern Construction and Recall of Designs, while the autistic sample showed a 3.3 point difference on these two spatial subtests. These *different patterns* may explain why the normative model did not fit the ASD data to achieve strong measurement invariance. Consequently, the spatial score *does not hold the same meaning* for children from the

ASD and normative samples. For autistic children, the spatial subtests may be tapping different abilities.

The failed measurement invariance is not attributable to group mean differences. As expected, the ASD group showed average lower scores on every subtest, and every factor. Clinicians administering the DAS-II to autistic children might consider placing more emphasis on the verbal and nonverbal reasoning composite scores instead of the spatial or composite GCA (General Conceptual Ability). Historically, some ASD clinicians and researchers have relied upon the SNC (Special Nonverbal Composite) instead of the GCA because the SNC excludes the verbal composite. The logic is that verbal subtests may be poor indicators of intelligence of an autistic person, given the communication difficulties inherent in the diagnosis. However, our results suggest that the spatial score, not the verbal score, poses validity issues. We suggest that clinicians avoid interpreting the SNC and GCA and instead defer to the verbal and nonverbal reasoning standardized scores when utilizing the DAS-II. For example, an autistic child with a true spatial intelligence of 95 could record a DAS-II spatial composite score of 92, or 98; their true spatial intelligence could be over- or under-estimated, depending on the pattern of their Pattern Construction and Recall of Designs subtest scores. Since it is not possible at this time to quantify and predict how each autistic child's true spatial ability would be misrepresented by the DAS-II Spatial Composite score, we recommend avoiding interpretation of the DAS-II Spatial composite score for autistic children, and consequently their SNC and GCA scores.

Implications for 'IQ splits' in ASD

These results suggest that the oft discussed autistic verbal-nonverbal 'IQ splits' are likely to be real, and not an artifact of the DAS-II functioning differently in autistic children than normative sample children. The ASD IQ splits refer to differences between the verbal and nonverbal reasoning scores and do not include the spatial score. Even when such studies of IQ splits have used the DAS-II, such as Nowell and colleagues' (2015) investigation of splits in the present ASD dataset, the authors analyzed only the verbal and nonverbal composite scores, not the spatial composite score. The verbal and nonverbal composite scores reflect true differences in verbal and nonverbal abilities, according to the partial scalar invariance achieved in the present analysis.

Issues with modeling the nonverbal reasoning factor

Surprisingly, with the six core subtests we were unable to fit properly the higher-order factor model that the publisher emphasizes. The published documentation only provides higher-order model results for the infrequently used full battery of 6 core and 6 diagnostic subtests. The problem in fitting the higher-order model to the 6 core subtests lay in the nonverbal factor loading entirely onto the general factor and providing no specific variance. This issue resurfaced when we attempted to fit a 3-factor bifactor model, which differs from the higher-order model in that the general factor is orthogonal to the group factors and not permitted to correlate with them. Both nonverbal subtests loaded directly onto *g*, not the nonverbal factor. The issues were even more salient in the ASD dataset, where the nonverbal factor showed an even higher and more improbable loading (1.045) onto the general factor in the higher-order model, and the 3-factor bifactor model failed to converge at all. Thus, the issue of the nonverbal factor not

existing independently of *g* seems intrinsic to the DAS-II and not specific to a particular dataset. Eliminating either the general factor (correlated 3-factor model) or the nonverbal factor (2-factor bifactor model) resolved the convergence issue and the resulting models showed excellent fit. We are not the first to report that the nonverbal factor may be absorbed entirely by the general factor (Dombrowski et al., 2018), and that second-order factors may provide little additional specific variance over and above *g* (Canivez & McGill, 2016; Dombrowski et al., 2019). However, it merits mention that when additional DAS-II subtests enter into the model, such as all 20 subtests, other groups have replicated the publishers' reported higher-order model (Dombrowski et al., 2019; Keith, Low, Reynolds, Patel, & Ridley, 2010).

Limitations

The primary limitation of this study concerns the depth at which we can understand the bias. The partial invariance methods used here allow us to identify which factor(s) shows bias, and the directionality of the bias for each subtest. We cannot, however, transform differences in intercepts to differences in DAS-II subtest points and suggest a correction. We also do not know why the bias occurs in these particular subtests. Future research to answer these questions would involve an item-level analysis of differential item functioning between the normative and autistic samples.

A second limitation concerns the missing verbal subtest data in the ASD dataset, which was systematically missing for individuals with lower verbal abilities on other auxiliary verbal variables. Much autism research excludes individuals with low verbal abilities (Russell et al., 2019), and we wanted our results to generalize to this very

understudied population. Thus we included individuals with missing verbal data in the analyses, and the missing data may have affected model fit. To address this limitation, we reran all invariance analyses twice: with complete cases only and with imputed missing data for these subtests. In both alternative analyses, we found no meaningful change in results.

Finally, the SSC autistic sample used in this analysis, while large and diverse in terms of race and ethnicity, includes only simplex individuals, meaning individuals with no first degree relatives with ASD. If simplex ASD is found to be qualitatively different than multiplex ASD (where ASD is present in one or more first degree relatives), then these results may not generalize to multiplex ASD. At present, this limitation does not cause concern because no studies have identified significant differences in the pattern of cognitive abilities between simplex and multiplex ASD, to our knowledge.

Future directions

Measurement invariance for autistic individuals has not been investigated in other IQ assessments, such as the Wechsler or Stanford Binet scales, to our knowledge. Our DAS-II findings suggest that such future analyses may be important. Furthermore, future studies might test measurement bias in commonly used ASD measures by sex as larger datasets of females with ASD become available; measurement invariance can be detected with as few as 200 participants per group (Finch & French, 2016). Finally, DAS autistic norms could be developed to improve interpretability of the spatial subtest scores for autistic populations.

Conclusions

The DAS-II Spatial standardized score should be interpreted with caution for autistic children. This score likely includes measurement bias or artifacts present for autistic children that are absent in the normative sample children. The verbal and nonverbal reasoning standardized scores do hold the same meaning for both autistic and neurotypical children, according to these results from the largest samples analyzed to date.

References

Canivez, G. L., & McGill, R. J. (2016). Factor structure of the Differential Ability Scales–Second Edition: Exploratory and hierarchical factor analyses with the core subtests. *Psychological Assessment, 28*(11), 1475.

Charman, T., Pickles, A., Simonoff, E., Chandler, S., Loucas, T., & Baird, G. (2011). IQ in children with autism spectrum disorders: data from the Special Needs and Autism Project (SNAP). *Psychological Medicine, 41*(3), 619-627.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464-504.

Diamantopoulos, A., & Siguaw, J. A. (2003). *Introducing LISREL: A guide for the uninitiated*. Thousand Oaks, CA: Sage.

Dombrowski, S. C., Golay, P., McGill, R. J., & Canivez, G. L. (2018). Investigating the theoretical structure of the DAS-II core battery at school age using Bayesian structural equation modeling. *Psychology in the Schools*, *55*(2), 190-207.

Dombrowski, S. C., McGill, R. J., Canivez, G. L., & Peterson, C. H. (2019). Investigating the theoretical structure of the Differential Ability Scales—Second Edition through hierarchical exploratory factor analysis. *Journal of Psychoeducational Assessment*, *37*(1), 91-104.

Elliott, C. D. (2007a). Differential Ability Scales (2nd ed.). San Antonio, TX: Harcourt Assessment.

Elliott, C. D. (2007b). Differential Ability Scales (2nd ed.): Introductory and technical handbook. San Antonio, TX: Harcourt Assessment.

Finch, W. H., & French, B. F. (2016). Quantifying the influence of partial scalar invariance on mean comparisons: two proposed effect sizes. *International Journal of Quantitative Research in Education, 3*(4), 292-313.

Fischbach, G. D., & Lord, C. (2010). The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron, 68*(2), 192-195.

Goldstein, G., Allen, D. N., Minshew, N. J., Williams, D. L., Volkmar, F., Klin, A., & Schultz, R. T. (2008). The structure of intelligence in children and adults with high functioning autism. *Neuropsychology, 22*(3), 301.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*, 183-202.

Jorgensen, T. D., Kite, B. A., Chen, P. Y., & Short, S. D. (2018). Permutation randomization methods for testing measurement equivalence and detecting differential item functioning in multiple-group confirmatory factor analysis. *Psychological Methods, 23*(4), 708.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2019). semTools: Useful tools for structural equation modeling. R package version 0.5-2. Retrieved from https://CRAN.R-project.org/package=semTools

Keith, T. Z., Low, J. A., Reynolds, M. R., Patel, P. G., & Ridley, K. P. (2010). Higher-order factor structure of the Differential Ability Scales–II: Consistency across ages 4 to 17. *Psychology in the Schools, 47*(7), 676-697.

Muthén, L.K. & Muthén, B.O. (1998-2017). Mplus User's Guide. Eighth Edition. Los Angeles, CA: Muthén & Muthén.

Newman, D. A. (2014). Missing data: Five practical guidelines. *Organizational Research Methods*, *17*, 372-411.

Nowell, K. P., Schanding, G. T., Kanne, S. M., & Goin-Kochel, R. P. (2015). Cognitive profiles in youth with Autism Spectrum Disorder: An investigation of base rate discrepancies using the Differential Ability Scales—Second Edition. *Journal of Autism and Developmental Disorders*, *45*(7), 1978-1988.

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology*, *96*(5), 966-980.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Reynolds, C. R., & Lowe, P. A. (2009). The problem of bias in psychological assessment. In T. B. Gutkin & C. R. Reynolds (Eds.), *Handbook of school psychology* (4th ed., pp. 332-374). Hoboken, NJ: Wiley.

Rosseel Y (2012). "lavaan: An R Package for Structural Equation Modeling." *Journal of Statistical Software*, *48*(2), 1–36. http://www.jstatsoft.org/v48/i02/

Russell, G., Mandy, W., Elliott, D., White, R., Pittwood, T., & Ford, T. (2019). Selection bias on intellectual ability in autism research: A cross-sectional review and meta-analysis. *Molecular Autism, 10*(1), 9.

Siegel, D. J., Minshew, N. J., & Goldstein, G. (1996). Wechsler IQ profiles in diagnosis of high-functioning autism. *Journal of Autism and Developmental Disorders*, *26*(4), 389-406.

Thaler, N. S., Barchard, K. A., Parke, E., Jones, W. P., Etcoff, L. M., & Allen, D. N. (2015). Factor structure of the Wechsler Intelligence Scale for Children: Fourth Edition in children with ADHD. *Journal of Attention Disorders*, *19*(12), 1013-1021.

Thompson, M. S. (2016). Assessing measurement invariance of scales using multiple-group structural equation modeling. In K. Schweizer & C. DiStefano (Eds.), Principles and methods of test construction: Standards and recent advances (pp. 218-244). Boston, MA: Hogrefe.

Trundt, K. M., Keith, T. Z., Caemmerer, J. M., & Smith, L. V. (2018). Testing for construct bias in the Differential Ability Scales: A comparison among African American, Asian, Hispanic, and Caucasian children. *Journal of Psychoeducational Assessment*, *36*(7), 670-683.

van Buuren S, Groothuis-Oudshoorn K (2011). "mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software*, *45*(3), 1-67. https://www.jstatsoft.org/v45/i03/

Wicherts, J. M. (2016). The importance of measurement invariance in neurocognitive ability testing. *The Clinical Neuropsychologist*, *30*(7), 1006-1016.

Tables

Table 1.

Participant demographics

	Normative sample	**ASD sample**	**Normative sample with complete data**	**ASD sample with complete data**
N	2400	1316	2388	1197
% male [a]	50.0	87.4	50.0	87.9
Age in years, mean[SD]	12.0[3.5]	10.5[3.7]	12.0[3.5]	10.5[3.7]
DAS-II Global Composite Ability	99.9[15.2]	94.3[19.8]	99.9[15.2]	94.4[19.7]
DAS-II Nonverbal Composite	99.8[14.8]	93.4[19.1]	99.8[14.8]	95.0[18.6]
DAS-II Verbal Composite	100.0[15.1]	92.9[22.6]	100.0[15.1]	93.0[22.5]
DAS-II Spatial Composite	99.8[14.9]	95.1[18.2]	99.9[14.9]	96.3[18.1]
Subtest				
Matrices (n)	50.2[10.2]	46.7[12.3]	50.2[10.2]	47.5[12.1]
Pattern construction (s)	50.0[9.9]	48.9[11.6]	50.0[9.9]	49.6[11.7]
Recall of designs (s)	50.0[9.9]	45.6[11.6]	50.0[9.9]	46.4[11.4]
Seq. & quant. Reasoning (n)	50.2[10.3]	45.8[12.9]	50.2[10.3]	46.9[12.6]

Verbal similarities (v)	50.2[9.9]	46.2[13.9]	50.2[9.9]	46.3[13.8]
Word definitions (v)	50.1[9.8]	45.5[14.7]	50.1[9.8]	45.5[14.7]

Note. All DAS-II values show mean [SD] of the standard score; Lowercase letters in parentheses denote composite in which subtest is scored.

[a] Missing for 56 individuals with ASD

Table 2.

DAS-II subtest correlations

	Matrices	Pattern construc-tion	Recall of designs	Sequential and quant. reasoning	Verbal similarities	Word definitions
Matrices	1	0.54	0.48	0.62	0.50	0.49
Pattern construction	0.62	1	0.55	0.58	0.46	0.44
Recall of designs	0.58	0.67	1	0.51	0.43	0.42
Seq. & quant. reasoning	0.72	0.63	0.57	1	0.54	0.54
Verbal similarities	0.58	0.46	0.45	0.62	1	0.65
Word definitions	0.57	0.47	0.46	0.64	0.79	1

Note. The upper set of correlations depicts the normative dataset; the lower set depicts the ASD dataset.

Table 3.

Model fit statistics

Model	df	χ^2	p	CFI	TLI	RMSEA (90%)	SRMR	AIC
Unidimensional								
Normative	9	385.0	0.000	0.934	0.890	0.132 (0.121-0.143)	0.040	101,304.5
ASD-SSC	9	537.6	0.000	0.864	0.774	0.211 (0.196-0.227)	0.060	56,529.0
Correlated 3-factor (Oblique)								
Normative	6	4.9	0. 555	1.000	1.000	0.000 (0.000-0.024)	0.005	100,913.6
ASD-SSC	6	13.3	0.038	0.998	0.995	0.030 (0.007-0.053)	0.009	55,941.3
Higher-Order								
Normative	6	4.9	0.555	1.000	1.000	0.000 (0.000-0.024)	0.005	100,913.6
ASD-SSC	6	13.3	0.038	0.998	0.995	0.030 (0.007-0.053)	0.009	55,941.3
Bifactor, 3 factor								
Normative[a]	9	23.3	0.001	0.997	0.996	0.026 (0.013-0.039)	0.042	100,518.0

ASD-SSC	9	--	--	--	--	--	--	--
Bifactor, 2 factor								
Normative[a]	7	5.0	0.654	1.000	1.001	0.000 (0.000-0.020)	0.005	100,911.8
ASD-SSC	7	22.4	0.002	0.996	0.992	0.041 (0.023-0.060)	0.013	55,949.8
Measurement Invariance								
Configural	12	18.8	0.094	0.999	0.998	0.017 (0.000-0.032)	0.007	156,854.9
Metric	15	24.7	0.054	0.999	0.998	0.019 (0.000-0.031)	0.016	156,856.0
Scalar	18	133.7	<0.001	0.988	0.980	0.059 (0.050-0.068)	0.035	156,968.7
Partial scalar: Spatial[b]	17	29.6	0.030	0.999	0.998	0.020 (0.006-0.032)	0.019	156,857.3
Partial scalar: Nonverbal[b]	17	130.9	<0.001	0.988	0.979	0.060 (0.051-0.070)	0.035	156,967.8
Partial scalar: Verbal[b]	17	131.3	<0.001	0.988	0.979	0.060 (0.051-0.070)	0.033	156,968.3

[a] Results from n=2388 with participants with missingness excluded; model did not converge with dataset with missing data (n=2400)

[b] The intercept of one subtest on the respective spatial, nonverbal, or verbal factor is free to vary between groups; model fit is identical in the two models of the factor's two subtests varying

Table 4.

Unstandardized intercepts and means, by model

	Configural	Metric	Scalar	Partial scalar[a]	Partial scalar[b]	Partial scalar, strict[a]
Factor means	Fixed at 0	Fixed at 0	Free	Free	Free	Free
Factor loadings	Free	Invariant	Invariant	Invariant	Invariant	Invariant
Factor intercepts	Free	Free	Invariant	5/6 invariant	5/6 invariant	5/6 invariant
Residuals	Free	Free	Free	Free	Free	Invariant
Normative						
Verbal	0	0	0	0	0	0
Nonverbal	0	0	0	0	0	0
Spatial	0	0	0	0	0	0

Matrices (n)	50.2	50.2	50.3	50.3	50.3	50.3
Pattern construction (s)	50.0	50.0	50.5	50.0	50.0	50.0
Recall of designs (s)	50.0	50.0	49.3	50.0	50.0	50.0
Seq. & quant. Reasoning (n)	50.2	50.2	50.1	50.1	50.1	50.1
Verbal similarities (v)	50.2	50.2	50.3	50.3	50.3	50.3
Word definitions (v)	50.1	50.1	50.0	50.0	50.0	50.0
ASD						
Verbal	0	0	-0.64	-0.64	-0.64	-0.65
Nonverbal	0	0	-0.50	-0.50	-0.50	-0.51
Spatial	0	0	-0.34	-0.14	-0.63	-0.14
Matrices (n)	46.6	46.6	50.3	50.3	50.3	50.3

Pattern construction (s)	48.9	48.9	50.5	50.0	53.8	50.0
Recall of designs (s)	45.6	45.6	49.3	46.6	50.0	46.6
Seq. & quant. Reasoning (n)	45.7	45.7	50.1	50.1	50.1	50.1
Verbal similarities (v)	45.5	45.4	50.3	50.3	50.3	50.3
Word definitions (v)	44.6	44.6	50.0	50.0	50.0	50.0

Note. lowercase letters denote factor onto which subtest loads. See Table S4 for unstandardized intercepts and means for other partial scalar invariance models.

[a]The Recall of designs intercept was freed to vary between groups

[b]The Pattern construction intercept was freed to vary between groups

Table 5.

Models compared with permutation testing on multiple fit indices

	χ^2	CFI	RMSEA	TLI	AIC	SRMR
Model Comparison						
Configural vs baseline						
Delta	20.2	0.999	0.019	0.998	156,854.9	0.006
p value	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999
Metric vs configural						
Delta	7.1	<0.001	0.002	<0.001	1.09	0.005
p value	0.13	0.12	0.048	0.064	0.13	0.099
Scalar[a] *vs metric*						
Delta	118.7	-0.011	0.041	-0.017	112.7	0.013
p value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Partial scalar[a] *vs metric*						
Delta	5.3	0.000	0.001	0.000	1.3	0.001
p value	0.069	0.066	0.048	0.051	0.069	0.016

Partial scalar[a] *vs scalar*[a]						
Delta	-113.4	0.010	-0.040	0.017	-111.4	-0.012
p value	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999
Partial scalar[a] *vs strict partial scalar*[a]						
Delta	29.1	-0.002	0.008	-0.002	17.1	0.003
p value	<0.001	<0.001	<0.001	<0.001	<0.001	0.1

[a]The Recall of designs intercept was freed to vary between groups

Note. More complex model being tested appears first. Permutation testing executed using the permuteMeasEq function in the semTools R package.

Figures

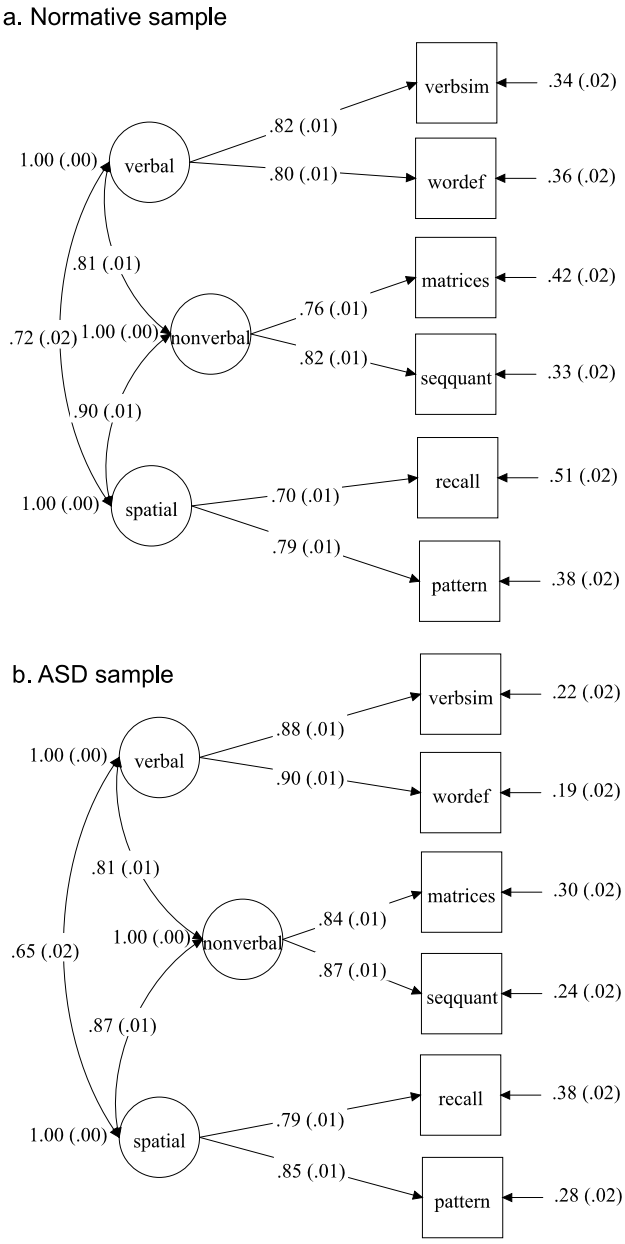

Figure 1. Correlated three-factor model for normative and ASD samples. Abbreviations: VerbSim: Verbal similarities; WordDef: Word Definitions; Pattern: Pattern Construction; Recall: Recall of Designs; SeqQuant: Sequential and Quantitative Reasoning

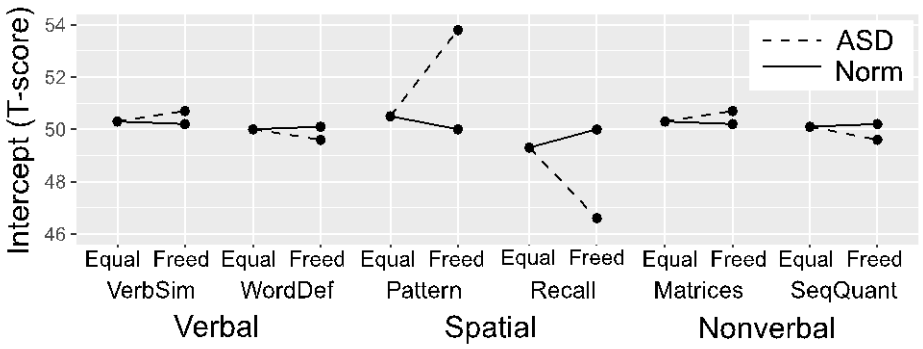

Figure 2. Change in intercept when no longer constrained equal between groups. Abbreviations: Cons: Constrained; VerbSim: Verbal similarities; WordDef: Word Definitions; Pattern: Pattern Construction; Recall: Recall of Designs; SeqQuant: Sequential and Quantitative Reasoning

APPENDIX

Chapter 1

Table S1.

Descriptive characteristics of participants included in neuropsychiatric questionnaires: Social Communication Questionnaire, Lifetime

Region	N	Age mean(sd)	Age range	% Male
AB/AC del group	13	14.5 (10.4)	6-38	46%
A-B Deletion	10	15.6 (12.9)	6-38	50%
A-C Deletion	3	12.7 (6.3)	6-18	33%
BD/CD del group	12	9.5 (5.2)	4-18	50%
B-D Deletion	8	10 (3.5)	6-14	38%
C-D Deletion	4	8.9 (7.8)	4-18	75%
Classic AD del	70	8.7 (3.4)	4-17	77%
BD/CD dup group	9	7.8 (2.8)	5-11	56%
B-D Duplication	7	7.7 (2.2)	6-9	57%
C-D Duplication	2	7.9 (4.3)	5-11	50%
Classic AD dup	29	8.4 (3.2)	4-14	76%
ASD	70	7.8 (3.3)	3-14	80%
TDC	73	7.8 (3.5)	2-14	77%

Table S2.

Descriptive characteristics of participants included in neuropsychiatric questionnaires: Social Responsiveness Scale-2

Region	N	Age mean(sd)	Age range	% Male
AB/AC del group	13	12.4 (10.3)	2-38	46%
A-B Deletion	10	12.3 (12.1)	2-38	50%
A-C Deletion	3	12.7 (6.3)	6-18	33%
BD/CD del group	12	16.2 (14.8)	4-42	50%
B-D Deletion	8	15.8 (14.9)	6-42	38%
C-D Deletion	4	16.7 (16.8)	4-40	75%
Classic AD del	70	7.7 (3.8)	2-16	77%
BD/CD dup group	9	6.2 (3.4)	2-11	56%
B-D Duplication	7	5.3 (3.1)	2-9	57%
C-D Duplication	2	7.9 (4.3)	5-11	50%
Classic AD dup	29	7.4 (3.5)	3-14	76%
ASD	70	7.8 (3.3)	3-14	80%
TDC	73	7.8 (3.5)	2-14	77%

Table S3.

Descriptive characteristics of participants included in neuropsychiatric questionnaires: Vineland Adaptive Behavior Scales-II

Region	N	Age mean(sd)	Age range	% Male
AB/AC del group	13	11.6 (10.4)	2-38	46%
A-B Deletion	10	11.5 (12.3)	2-38	50%
A-C Deletion	3	11.7 (5.3)	6-15	33%
BD/CD del group	12	3 (2.9)	0-8	50%
B-D Deletion	8	4.1 (3.3)	2-8	38%
C-D Deletion	4	1.3 (1.3)	0-2	75%
Classic AD del	70	7.3 (4)	2-16	77%
BD/CD dup group	9	5.7 (2.7)	2-9	56%
B-D Duplication	7	5.3 (3.1)	2-9	57%
C-D Duplication	2	6.4 (2.2)	5-8	50%
Classic AD dup	29	7.1 (3.4)	3-14	76%
ASD	70	7.8 (3.3)	3-14	80%
TDC	73	7.8 (3.5)	2-14	77%

Table S4. Descriptive characteristics of participants included in neuropsychiatric questionnaires: Child and Adolescent Symptom Inventory, 4th Edition, Revised

Region	N	Age mean(sd)	Age range	% Male
AB/AC del group	13	11.6 (10.4)	2-38	46%
A-B Deletion	10	11.5 (12.3)	2-38	50%
A-C Deletion	3	11.7 (5.3)	6-15	33%
BD/CD del group	12	3 (2.9)	0-8	50%
B-D Deletion	8	4.1 (3.3)	2-8	38%
C-D Deletion	4	1.3 (1.3)	0-2	75%
Classic AD del	70	7.3 (4)	2-16	77%
BD/CD dup group	9	5.7 (2.7)	2-9	56%
B-D Duplication	7	5.3 (3.1)	2-9	57%
C-D Duplication	2	6.4 (2.2)	5-8	50%
Classic AD dup	29	7.1 (3.4)	3-14	76%
ASD	70	7.8 (3.3)	3-14	80%
TDC	73	7.8 (3.5)	2-14	77%

Note. Abbreviations: ASD: autism spectrum disorder; d; Cohen's d effect size; del: typical 22q11.2 Deletion Syndrome involving LCR-A to D, dup: typical 22q11.2 Duplication Syndrome involving LCR-A to D, TDC: typically developing controls; SCQ: Social Communication Questionnaire, Lifetime; SRS: Social Responsiveness Scale

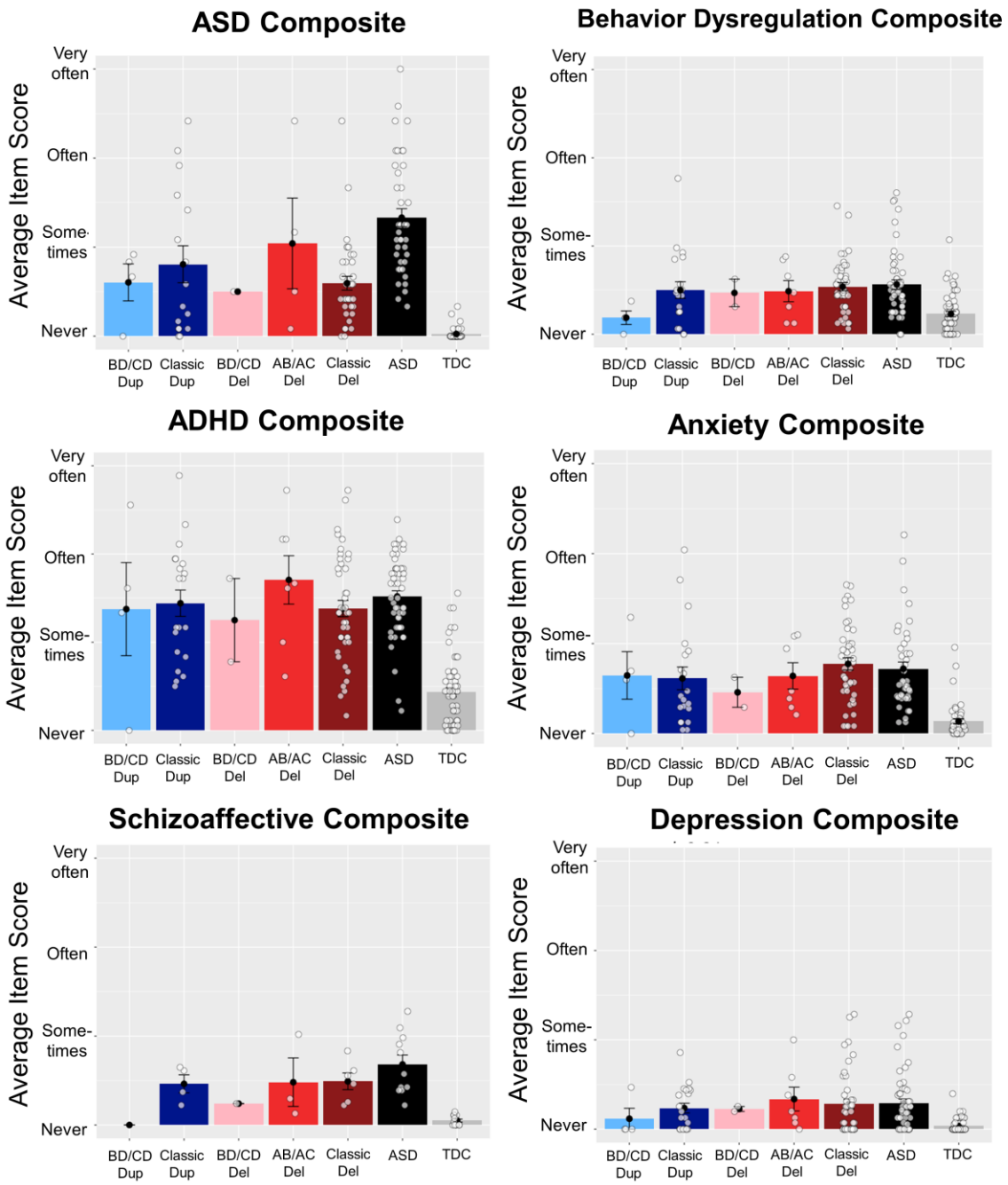

Figure S1. Patterns in parent-reported psychiatric symptoms across individuals with classic or nested 22q11.2 duplications or deletions compared to typically developing controls. Group means and standard errors on six composite indices of the CASI-4R, a parent-report measure of psychiatric symptoms in DSM-5 diagnoses. Groups include the

"BD/CD" duplication (light blue) or deletion (pink) groups (individuals with nested duplication or deletion involving LCR-B to LCR-C or D), the "AB/AC" deletion group in red (individuals with nested deletion of LCR-A to B or C), the "Classic Del" group in dark red (individuals with typical deletion spanning LCR-A to LCR-D), "Classic Dup" group in dark blue (individuals with typical duplication spanning LCR-A to LCR-D), "ASD" group in black (individuals with non-syndromic autism spectrum disorder), and "TDC" group in gray (typically developing children). Higher scores on the CASI-4R indicate higher symptom levels. The "BD/CD" deletion (pink) and duplication (light blue) groups showed similar or lower levels of symptoms compared to the other deletion or duplication groups, respectively (see table 4 for details). All 22q11.2 groups show higher symptom levels than the typically developing controls. Abbreviations: ASD: autism spectrum disorder; CASI-4R: Child and Adolescent Symptom Inventory-4R; del: classic 22q11.2 Deletion Syndrome involving LCR-A to D, dup: classic 22q11.2 Duplication Syndrome involving LCR-A to D, TDC: typically developing controls

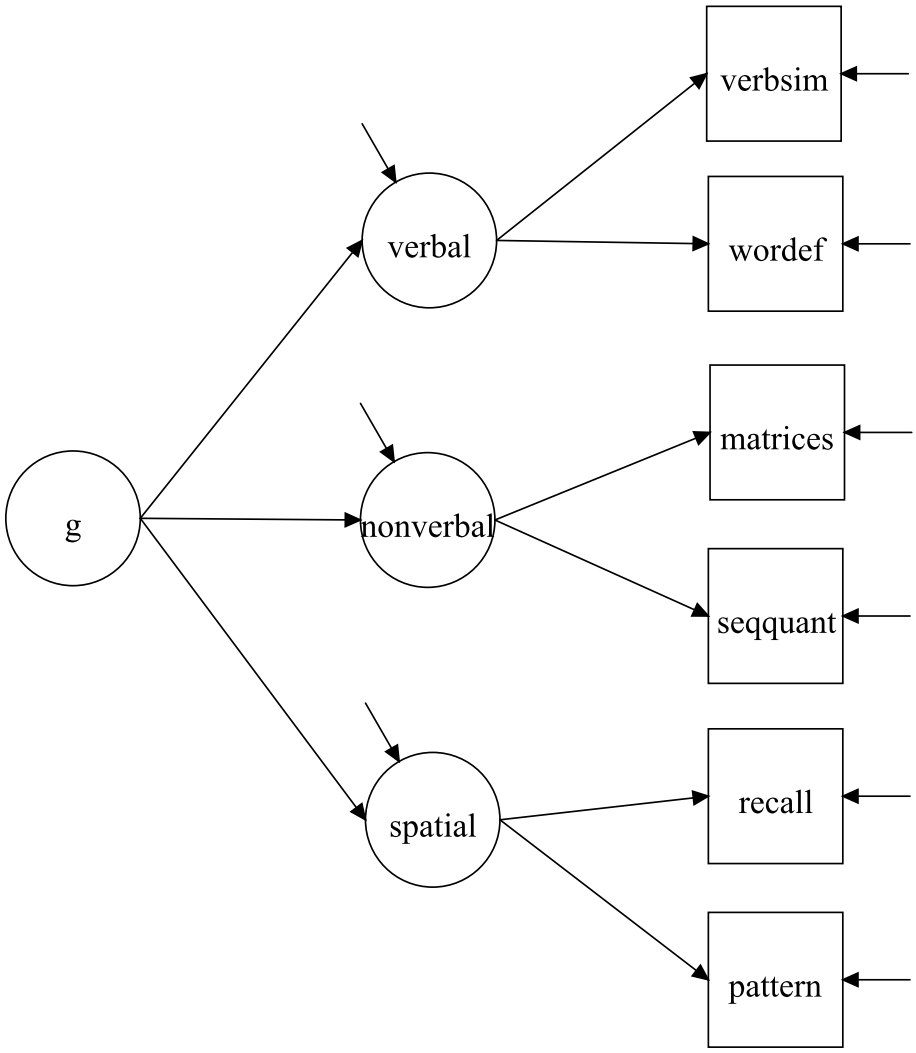

Figure S1. Higher-order model and three-factor bifactor model. In the higher-order model, the three specific factors (verbal, nonverbal, and spatial) load onto one higher order, general factor *g.* In the three-factor bifactor model, the model is similar but the general factor is assumed to be orthogonal to the specific factors, and the correlations between specific factors can be estimated. We fixed the between-factor correlations at 0 to permit model identification (see Table S1).

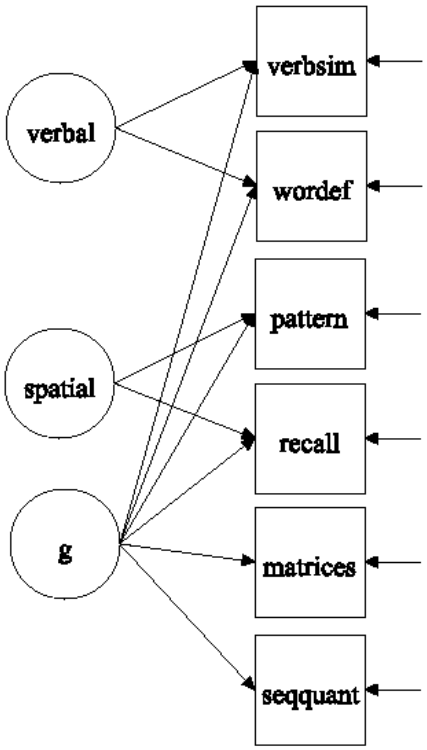

Figure S2. Two-factor bifactor model. The nonverbal factor is omitted and the nonverbal subtests (Recall of Designs and Pattern Completion) load directly on the general factor *g*.

Table S1.

Standardized factor loadings and correlations for each model

	Correlated 3-Factor (ie Oblique)	Higher-order	3-factor Bifactor	2-factor Bifactor[a]
Normative				
Correlations				
Nonverbal-verbal	0.81	Not estimated	0	0
Spatial-verbal	0.72	Not estimated	0	0
Spatial-nonverbal	0.90	Not estimated	0	0
Loadings				
Verbal similarities (v)	0.82	0.82	0.81	0.47; 0.66
Word definitions (v)	0.80	0.80	0.81	0.47; 0.65
Matrices (n)	0.76	0.76	0.78	n/a; 0.76
Seq. & quant. Reasoning (n)	0.82	0.82	0.80	n/a; 0.82
Recall of designs (s)	0.70	0.70	0.76	0.33; 0.63
Pattern construction (s)	0.79	0.79	0.73	0.33; 0.70
Verbal (*g*)	no *g* factor	0.81	0.81	Not estimated

Nonverbal reasoning (*g*)	no *g* factor	1.01	1.00	Not estimated
Spatial (*g*)	no *g* factor	0.89	0.90	Not estimated
ASD				
Correlations				
Nonverbal-verbal	0.81	Not estimated	--	0
Spatial-verbal	0.65	Not estimated	--	0
Spatial-nonverbal	0.87	Not estimated	--	0
Loadings				
Verbal similarities (v)	0.88	0.88	--	0.56; 0.70
Word definitions (v)	0.90	0.90	--	0.53; 0.72
Matrices (n)	0.84	0.84	--	n/a; 0.84
Seq. & quant. Reasoning (n)	0.87	0.87	--	n/a; 0.88
Recall of designs (s)	0.79	0.79	--	0.42; 0.68
Pattern construction (s)	0.85	0.85	--	0.42; 0.73
Verbal (*g*)	no *g* factor	0.78	--	Not estimated
Nonverbal reasoning (*g*)	no *g* factor	1.05	--	Not estimated
Spatial (*g*)	no *g* factor	0.83	--	Not estimated

Note. Lowercase letters in parentheses denote factor onto which subtest loads.

[a] Loading on the specific factor are noted first, followed by a semi-colon and loading on the general factor *g*

Table S2.

Standardized factor loadings and correlations, for measurement invariance models

	Configural	Metric	Scalar	Partial scalar[a]
Factor means	Fixed at 0	Fixed at 0	Free	Free
Factor loadings	Free	Invariant	Invariant	Invariant
Factor intercepts	Free	Free	Invariant	5 of 6 invariant
Normative				
Correlations				
Nonverbal-verbal	0.81	0.81	0.81	0.81
Spatial-verbal	0.72	0.72	0.73	0.72
Spatial-nonverbal	0.90	0.90	0.90	0.90
Loadings				
Verbal similarities (v)	0.82	0.80	0.80	0.80
Word definitions (v)	0.80	0.82	0.82	0.82
Matrices (n)	0.76	0.76	0.76	0.76
Seq. & quant. Reasoning (n)	0.82	0.82	0.82	0.82
Recall of designs (s)	0.70	0.71	0.72	0.71

Pattern construction (s)	0.79	0.78	0.77	0.78
ASD				
Correlations				
Nonverbal-verbal	0.81	0.81	0.81	0.81
Spatial-verbal	0.65	0.65	0.65	0.65
Spatial-nonverbal	0.87	0.87	0.88	0.87
Loadings				
Verbal similarities (v)	0.88	0.90	0.89	0.89
Word definitions (v)	0.90	0.89	0.89	0.89
Matrices (n)	0.84	0.84	0.84	0.84
Seq. & quant. Reasoning (n)	0.87	0.87	0.87	0.87
Recall of designs (s)	0.79	0.78	0.79	0.78
Pattern construction (s)	0.85	0.85	0.83	0.85

Note. lowercase letters in parentheses denote factor onto which subtest loads

[a]The Recall of designs intercept was freed to vary between groups

Table S3.

Unstandardized factor means and subtest intercepts for partial scalar invariance models

	Matrices	Seq. & quant. reasoning	Verbal similarities	Word definitions
Normative				
Verbal	0	0	0	0
Nonverbal	0	0	0	0
Spatial	0	0	0	0
Matrices (n)	50.2	50.2	50.3	50.3
Pattern construction (s)	50.5	50.5	50.5	50.5
Recall of designs (s)	49.3	49.3	49.3	49.3
Seq. & quant. reasoning (n)	50.2	50.2	50.1	50.1
Verbal similarities (v)	50.3	50.3	50.2	50.2
Word definitions (v)	50.0	50.0	50.1	50.1
ASD				
Verbal	-0.64	-0.64	-0.67	-0.61
Nonverbal	-0.53	-0.46	-0.50	-0.50

Spatial	-0.34	-0.34	-0.34	-0.34
Matrices (n)	50.7	50.2	50.3	50.3
Pattern construction (s)	50.5	50.5	50.5	50.5
Recall of designs (s)	49.3	49.3	49.3	49.3
Seq. & quant. Reasoning (n)	50.2	49.6	50.1	50.1
Verbal similarities (v)	50.3	50.3	50.7	50.2
Word definitions (v)	50.0	50.0	50.1	49.6

Note. lowercase letters denote factor onto which subtest loads. See Table 4 for unstandardized intercepts and means for other measurement invariance models.