Assessing Early Child Development: Issues of Measurement Invariance and Psychometric Validity

Eric Kwame Duku

Thesis submitted to the

Faculty of Graduate and Postdoctoral Studies

In partial fulfilment of the requirements

For the Ph.D. degree in Education

Teaching, Learning and Evaluation Concentration

Faculty of Education

University of Ottawa

Declaration of academic achievement

Eric Duku, the author of the manuscript, "Investigating the measurement properties of the Social Responsiveness Scale in preschool children with Autism Spectrum Disorders" is the primary author of this article. As the primary author, contributions include: theoretical and methodological formulations for the research, including the research proposal, literature review, analyzing data, manuscript preparation and manuscript revision. The data used for this manuscript came from the Pathways in Autism Study. The investigators of the Pathways in Autism Study are co-authors of this manuscript including Dr. Tracy Vaillancourt who is the thesis supervisor. Dr. Vaillancourt offered input and expertise during each phase of the research formulation and manuscript preparation. As per guidelines for training and publication, the co-authors (or principal investigators) offered feedback and approval of the final and revised manuscripts for submission. This manuscript has been published in the *Journal of Autism and Developmental Disorders*.

Eric Duku is the primary author of the manuscript "Measurement equivalence of the autism symptom phenotype in children and youth". Eric Duku conceptualized the manuscript from the theoretical and methodological formulations from the research proposal, conducted the literature review, analyzed the data, and prepared the manuscript for submission. The second author and thesis supervisor, Dr. Vaillancourt, offered input and expertise during each phase of the research formulation and manuscript preparation. As per guidelines for publication, the coauthors offered input and approval of the final manuscript for submission. The manuscript has been submitted to the *Journal of Child Psychology and Psychiatry* and is currently in press.

Finally, Eric Duku is the primary author of the manuscript "Validation of the BRIEF-P in a sample of Canadian preschool children". Eric Duku conceptualized the manuscript from the theoretical and methodological formulations from the research proposal, conducted the literature review, analyzed the data, and prepared the manuscript for submission. The second author, Tracy Vaillancourt, is the thesis supervisor who offered input and expertise during each phase of the research formulation and manuscript preparation. The manuscript has been submitted to the *Journal of Child Neuropsychology* and is in press.

Abstract

The measurement of reliable and valid indicators of early child development is necessary for assessing phenomena and is useful in the monitoring of ongoing efforts to eradicate inequalities in the social determinants of health. There is an increasing awareness of the contextual, cultural, and developmental influences on constructs used in early child development (ECD) research. Using a measurement perspective, this dissertation examined the issue of measurement invariance and psychometric validity in early child development research. A construct violates the principle of invariance when two persons from different populations who are theoretically identical on the construct being measured have different scores on it.

This dissertation consists of three journal-style manuscripts (published or under review) that were used as examples to address the importance of the issue of measurement invariance and psychometric validity in ECD research using data from two unique areas: autism and executive functioning. The three data sets were collected on pre-school children with parents and or teachers as informants and were chosen to represent different levels of data collection – clinical, community, and population. These data sets allowed for the examination of measurement invariance by type of informant, sex, and age of child. The results from the three studies illustrate the importance of assessing measurement invariance in ECD and whether or not the instruments examined can be used to assess sub-group differences with confidence.

A lack of measurement invariance found for two of the studies, suggests that observed group differences in latent constructs could be attributed, in part, to measurement bias. More importantly, bias in the measurement of the constructs of severity of social impairment symptoms in autism, and executive functioning across groups could have an impact on services

such as patient treatment. These biases could also influence public policy development, particularly when there may be an underlying need for a cross-group approach where belief systems may affect the meaning and structure of constructs.

In summary, measurement invariance should be a prerequisite for making any meaningful comparisons across groups. A requirement of establishing measurement invariance should be included in the guidelines for comparative research studies as a necessary first step before an instrument is adopted for use.

Acknowledgements

I would never have been able to finish my dissertation without the guidance and support of my supervisor, committee members, friends and colleagues, and the love and support from my wife and family.

I would like to thank my supervisor, Professor Tracy Vaillancourt for her guidance, encouragement, unending support and academic collaboration throughout my academic journey. I would also like to thank the members of my committee, Professors David Smith, Dave Trumpower, David Smith and Brad Cousins for their thoughtful feedback and for pushing me beyond my comfort zone. Their suggestions and constructive feedback facilitated my journey. I am also grateful to my external examiner, Professor Todd Little for providing great insights on my dissertation and considerations for present and future research.

I am grateful for the financial support I received through University of Ottawa and the Canada Graduate Doctoral Scholarship from the Social Sciences and Humanities Research Council. The financial support was encouraging and instrumental in making this journey feasible. I should also note that this research would not have been possible without data and I would like to especially acknowledge the Pathways in ASD Study, the Autism Genome Project, and the Toddler Study for enabling access to the data used in my dissertation. I will also to take this opportunity to acknowledge my various co-authors who helped in the finalization of the manuscripts.

Next, I would also like to acknowledge my colleagues, Professors Magdalena Janus and Peter Szatmari, and members of the Offord Centre for Child Studies (too numerous to list) for your support and encouragement throughout my journey. To my friend and colleague, Stelios Georgiades, thank you for your mentorship, support and guidance as we both are going through

similar journeys. Efharisto, for your patience and wishing you the best as you get to end of your journey too and I look forward to our post dissertation outdoor meetings and coffee. To my fellow doctoral students and graduates (PiPsters), Amy Chen, Julie Comber, Osnat Fellus, Dr. Maria Gordon, Nathalie Gougeon, Dr. Joan Harrison, Christine Johnson, Jeela Jones, and Dr. Shari Orders, thank you for your support during this journey. We certainly have had fun, interesting and diverse experiences during our journeys. As always I send each of you positive energy as we continue on our various journeys. Thank you, Ms. Heather Brittain of the Peer Relations Lab, for your support and guidance in navigating challenges during my journey. To all the members of the Brain and Behaviour Lab, I am glad I to have been part of such an amazing group. I cannot forget my dear friend, Ms. Bev Martin, who over the last quarter century has also given me support, guidance and encouragement. It is hard to believe I am still the same shy person from years ago.

I would also like to acknowledge and thank my parents, Ms. Juliana Addo-Danquah (deceased) and Mr. Kofi Duku, most importantly for starting me on my life journey and for helping build the foundation for this journey. My thesis is dedicated to the memory of my mother.

Finally, I am indebted to my wife and family – Lisanne, Jairus, Alyssa and Jordan, for their love, phenomenal support, encouragement, and understanding, and for putting up with the amount of time and energy spent on my journey. Words cannot express how grateful I am to you. My journey would not have been possible without your support, understanding, and enlightenment along the way. I would also like to acknowledge the love and support of Arthur and Judith Kelsie during my journey. Last but not least, I would also like to acknowledge our

furry companions, Sassy, Kitka, Pepper, Babs, Kira, Isis, Blossom and Buster who have also provided comfort, support and life experiences during my studies.

Table of Contents

List of Figures

List of Tables

List of Appendices

Chapter1. Introduction

The pre-school period is a critical stage of human development and there is a strong relationship between early child development and concurrent and future well-being including healthy weight, better mental health, lower heart disease, higher competence in literacy and numeracy, lower criminality, and greater economic participation (Irwin, Siddiqi & Hertzman, 2008). Many constructs and indicators are available and are used to measure health disparities and inequalities in early child development (ECD) all over the world. A construct (or latent construct) is a concept that cannot be measured directly, but rather through observed variables. Ensuring the comparability of constructs when conducting comparative research and testing for between-group differences (as well as change over time) is of great importance in being able to monitor health disparities across the world. Of equal importance is establishing both comparable and equivalent constructs and items across diverse groups within countries for the purposes of ECD research. Indeed, adequate care needs to be taken to ensure that constructs developed in the context of one group (or culture or language) are not inappropriately applied to another group.

Guidelines of the International Testing Commission (ITC) for the fair testing and the use of instruments for comparative purposes state that for appropriate interpretation of the results of tests, it is essential to be confident in the ability of instruments and measures to tap into the same constructs in all settings (ITC, 2001; ITC, 2011). However, equivalence of constructs and items across diverse groups is rarely explicitly examined in ECD research and many researchers usually assume constructs are equivalent across groups without checking this assumption or infer equivalence using standard psychometric tests such as test-retest reliability and content or external validity (Knight & Zerr, 2010). Thus the guidelines for fair testing and adapting

instruments, although well documented, are not always adhered to in ECD research. A methodological approach that allows testing a hypothesis that an instrument is indeed applicable and fair across different groups is based on the concept of *measurement invariance* (van de Vijver & Tanzer, 2004).

The concept of measurement invariance (or *measurement equivalence* or *cross-cultural validity*) is a fundamental component of measurement and comparative research. By definition, a measure or scale is assumed to be invariant if, and when, members of different populations who have the same characteristics and attributes receive the same observed score (Teresi, 2006). Measurement invariance occurs when a construct shows the same psychometric properties across groups such as age, sex, ethnic group, type of sample, and respondent (Horn & McArdle, 1992; Raju, Laffitte, & Byrne, 2002; Taris, Bok, & Meijer, 1998; Vandenberg 2002). A central principle of measurement invariance is that the scores or ratings across groups must be on the same scale and that the empirical relationships between test items and traits of interest must remain stable across groups (Reise, Widaman, & Pugh, 1993). A construct violates the principle of invariance when two persons from different populations who are theoretically identical on the construct being measured have different scores on it (Schmitt & Kuljanin, 2008). For example, children are regularly compared on severity of autistic symptoms by sex using means and variances (Hoffman, 2009; Szatmari et al, 2012). Although these approaches are useful for answering specific questions, this is not the same as comparing the underlying measurement structure of the autism symptom phenotype (i.e., testing measurement equivalence) across sex (Vandenberg & Lance, 2000).

In ECD, there is a lot of work to be done to understand the factors that challenge the measurement invariance of constructs used in comparative studies and to ensure that instruments

used minimize bias as much as possible. Accordingly, it is important that ECD researchers examine and address the issue of measurement invariance to fill this gap within the ECD literature.

Literature Review

In this review of the literature, a general overview of the use of measurement invariance in comparative research is discussed. Since the comparative research field contains a wide variety of research, the main purpose of this section is to highlight available research that provides both background and context to this study. Within this general overview, various contextual factors and characteristics that could affect the measurement invariance of constructs are discussed. These factors include age, informant (or respondent), type of sample, and culture (or ethnicity or nationality). These characteristics feature prominently in early child development research, especially since data are usually obtained through an informant or through direct observation.

The history of measurement invariance is long, diverse, and multi-disciplinary. Although the concern for measurement invariance originally arose from the use of admission tests for colleges and universities (Meredith & Teresi, 2006), interest in dealing with the issues of measurement invariance is increasing, particularly with regard to the methodology used to make decisions regarding the appropriateness of comparisons across populations, time, and response mediums (Vandenberg, 2002). Establishing the invariance of a construct is a precondition for making substantive group comparisons (Vandenberg & Lance, 2000). There are circumstances that can threaten the quality of measurement tools, that are not directly addressable through approaches such as the calculation of reliability coefficients and assessment of psychometric properties, yet many organizations and researchers have approached this task using this type of

statistical approach (Boltë, Poutska, & Constantino, 2008; Paunonen & Ashton, 1998). It can be further argued that rather than testing for the presence of such circumstances, most researchers have simply ignored the potential presence of these conditions.

Vandenberg and Lance (2000) in a lengthy review of the literature on measurement equivalence examined the inconsistencies in testing measurement invariance using the confirmatory factor analysis framework. They reviewed both the theoretical and methodological literature on the process of testing for measurement equivalence and examined the assessment of measurement invariance across populations. Based on the inconsistencies in procedures used to test for measurement equivalence or invariance, the authors recommended an integrative and comprehensive hierarchical approach to testing and establishing measurement equivalence. This approach, based on conceptual and statistical grounds, involves conducting sequences of the eight primary tests of measurement invariance: invariant covariance matrices, configural invariance, metric invariance, scalar invariance, invariant unique item variances, invariant factor variances, invariant factor covariances, and invariant factor means.

Vandenberg (2002) re-examined the issue of measurement equivalence and addressed some of the challenges in the analytical procedures. He recommended using procedures such as item response theory (IRT), which examine the way specific items behave across groups as a way of evaluating measurement invariance.

Byrne and Watkins (2003) also examined the issue of measurement invariance across groups with a review of the literature on disparities in the conceptualization of instrument equivalence as well as acceptance of replicability of factorial structure across groups. The authors examined the earlier work by Byrne and Campbell (1999), which had described how evidence of non-invariance can exist even if replicability of factorial structure across groups is

established. They recommended exploring lack of invariance through an identification of item bias, using an analysis of variance (ANOVA) based approach proposed by van de Vijver and Leung (1997). Specifically, each item is examined separately for evidence of bias with the item score serving as the dependent variable and the group and total score levels serving as independent variables.

Following the review by Vandenberg and Lance (2000), there have been several developments in the field, including the use of comparative fit indices (CFI) for comparing nested models when testing for measurement invariance, the use of *item parcels,* and the impact of *partial invariance* on estimated group differences in reliability and means (Bandalos, 2002; Little, Cunningham, Shahar & Widaman, 2002). Item parcelling is the practice of summing the scores or averaging the scores of at least two items to reduce the number of indicators in the measurement or to remedy the problem of non-normality of items scores in structural equation modeling (Bandalos, 2002). The allowance for "*partial invariance*" was proposed by Byrne, Shavelson, and Muthén (1989) under the condition that some items could show lack of invariance, so long as the latent construct could be defined equally across groups (Steenkamp & Baumgartner, 1998).

Evaluating constructs across groups and subgroups within a population helps support their degree of generalizability (Prince, 2008). Also, constructs from instruments need to be replicable and generalizable beyond the group for which they were developed. An extensive study of the psychometric properties of the construct needs to be performed, and even more importantly, one must assess the impact of characteristics such as culture (ethnicity), informant, and type of sample used on measurement invariance.

Culture

In the past three decades, comparative research related to validating instruments across cultures has grown rapidly due to the advancement of scientific investigations across cultures on a global scale. The types of constructs used and the variety of cultural groups compared by researchers have become increasingly diverse. Although measurement invariance methodology is not new, in ECD research, Prince (2008) points out that true cross-cultural comparative research is not commonly seen and where it is evident, it is carried out under the auspices of the international agencies like the World Health Organization rather than scientific research communities. The majority of research investigations within single societies continue to assume homogeneity of the population even though researchers are aware that sub-groups within these societies, such as ethnicity, sex, and age groups are often quite heterogeneous. These sub-groups can differ from one another or from the overall population and the differences can contribute to measurement invariance.

It is therefore important to recognize that instruments used in ECD research need to be validated in different cultures so that changes in child development outcomes can be monitored and compared. Toward this aim, there needs to be an evaluation of equivalence between the original instruments and their adapted versions. This process should not be restricted to situations involving different countries and languages, but local and regional contexts within a single culture need attention as well (Reichenheim & Moraes, 2007). It should also be noted that, even though researchers are aware that societies are often quite heterogeneous with respect to subgroups, the majority of research investigations within single cultural societies continue to assume homogeneity of the population (Boltë et al., 2008; Byrne & Campbell, 1999). This assumption is problematic because sub-groups often differ from one another, or from the overall population, and hence these differences can contribute to measurement non-invariance

(Lenartowicz, & Roth, 2001; Muthén, 1989; Niles, 1999; Reichenheim & Moraes, 2007). This is especially true in field research with convenience samples of social, educational, or occupational sub-groups (Muthén, 1989).

Sample

Sampling bias can affect measurement invariance and usually occurs when samples used are not randomly drawn. Samples for studies are often chosen for convenience and not for appropriateness. Thus it is possible that in the sample drawn, certain members of the population are under- or over-represented (Rottig, 2009). It is also important to evaluate measurement invariance of constructs when samples are drawn from two populations because sampling bias can lead to poor replication of results and mapping of constructs (van de Vijver & Leung, 2000).

Another issue is the size of the sample used in studies, in particular the small sample size used. There is currently no consensus on the issue of the minimum sample size required. Some guidelines suggest 10 to 20 participants per parameter and others suggest that the sample size should depend on the power, hypothesis being tested, and the model complexity (Weston & Gore, 2006). However, it has been shown that sample size affects the chi-square statistic and the precision of the factor loading parameters in confirmatory factor analyses (Kline, 2011). Kline (2011) also encourages researchers to use larger sample sizes when testing complex models.

Informant

In ECD research, the informant is typically the person most knowledgeable (PMK) about the target child (i.e., the child being studied). Most often the PMK is the mother, caregiver or the child's teacher. Various authors have stressed the importance of understanding the basis of informant agreement and disagreement since it is important for assessing child psychopathology (Achenbach, 1993; Offord et al., 1996). Kramer et al. (2003) have shown that it is necessary to

have information from multiple perspectives or informants to assess children's health, temperament, impairment, or socioemotional behaviour in the absence of gold standard constructs. Even where multiple informants provide information on the same children, there is little or no agreement among them. One possible explanation for this discrepancy could be the difference in perspectives or biases of the informants or the context in which children are observed.

Furthermore, data on constructs collected on children during the pre-school years are usually obtained through parent and or teacher reports. If these constructs, such as child anxiety or other childhood disorders that are measured based on parent or teacher-report items, could be measured directly (i.e., through a blood test or other clearly objective means), many of the challenges associated with non-invariance could be dealt with (Mellenbergh, 1989; Meredith & Teresi, 2006). There is also the added challenge of how to integrate the data from different informants because of differences in context and low agreement in ratings between informants. It is therefore important that researchers understand the reasons for differences in informant perception and differences in measurement models by informant. Dirks et al. (2012), suggest that investigators examine this difference in informant reports as meaningful clinical information and move toward approaches that embrace contextual variability in children's behaviour.

Theoretical framework

There are three basic theoretical approaches to cross-cultural research – *absolutist*, *relativist,* and *universalist* (Berry, Poortinga, Segall & Dasen, 2002). The *absolutist* approach in general assumes basic processes are species wide (or biological), and culture has minimal impact on these processes and measures constructed for these processes. Constructs are thus thought to be invariant across cultures and standard instruments can measure the same construct across

different cultures. Berry et al. state that such an assertion should be supported by strong empirical and theoretical evidence (2002). Conversely, the *relativist* approach assumes that culture has a strong influence on the variability in all behaviour. It is therefore impossible to use the same instruments across cultures and in their stead only local instruments should be used. Methodologically, when the relativist approach is adopted, comparative studies are avoided because the assessments or constructs are ethnocentric.

For this dissertation, the *universalist* approach which is the middle ground between the absolutist and relativist approach was adopted and adapted for use in this comparative ECD research (Berry et al, 2002). Using the universalist approach as the theoretical framework for this dissertation means starting the research by being more open to the possibility that group differences could have significant and important differences in the way constructs are expressed or understood. This also means not making any prior assumptions that constructs will be the same across groups such as informants, sex and age. The assumptions made under the universalist approach are that (a) the basic processes in life are similar across all species and membership in differing groups will have an impact on variations in behaviour; (b) measurement of constructs will be difficult to achieve using a context-free definition; and (c) measurement of constructs will require instruments adapted to the context or the group. Lastly, the universalist approach makes no assumption about the dimensionality of a construct across groups nor does it make any assumptions about the interrelationships of the dimensions across groups. This is therefore an issue that needs to be examined and ascertained before instruments used in ECD research are adapted (Hu & Trandis, 1985). If this can be done, then a construct can be operationalized across groups and the same instrument can be used to gather information. There is therefore a need to examine the degree to which a construct has the same dimensional structure

and the same interrelationships between dimensions across groups in comparative research involving children especially. In summary, the main goal of using the universalist approach is to examine characteristics or aspects of constructs that are common across groups and to use those aspects to develop and measure the constructs under study in different groups in ECD research. Evanoff (2004) summarizes it aptly by asserting that using the universalist approach it may then be possible to come up with a set of norms which would apply across all groups of children.

The universalist approach fits well in supporting the guidelines of the ITC which will also provide additional context for this dissertation. There are twenty-two ITC (2010) guidelines dealing with issues of the control and regulation of access to test materials, guidelines on the fair and ethical use of tests, and requirements for training and specifying the competence of test users. The guidelines fall into four main categories: context, test development and adaptation, test administration, and documentation and score interpretation. The ITC guidelines recommend key competencies such as knowledge, skills, and other personal and contextual factors needed for responsible test use. In addition to these competencies, the guidelines address issues of professional and ethical standards in testing including; rights of the test taker and other parties involved in the testing process; choice and evaluation of alternative tests; and test administration, scoring, interpretation, report writing and feedback.

Research Objectives

Important methodological issues, such as measurement invariance, and cross-cultural (or cross-group) validity, should concern researchers in all fields of study when developing scales or using established scales for comparative research. There is an increasing awareness of the contextual, cultural, and developmental influences on constructs used in ECD research.

However, research within the ECD literature has only emerged within the last decade in addressing issues on measurement invariance of constructs.

The objectives of this dissertation were therefore to: (1) examine the issue of measurement invariance and psychometric validity in early child development research; and (2) demonstrate the importance of assessing the measurement invariance of constructs in early child development research (specifically research on children with autism, and evaluation of executive functioning in children), and illustrate the impact of how a failure to do so could lead to biases or questionable conclusions.

Studies

This dissertation consists of three studies that address the importance of the issue of measurement invariance and psychometric validity in ECD research with data from two unique areas: autism in diagnosed children and executive functioning in typically developing children. The three data sets were collected on pre-school children with parents and or teachers as informants and were chosen to represent different levels of data collection – clinical and community respectively. The data sets also permitted the examination of measurement invariance by age and sex of the child and type of informant. Specifically, the objectives of this dissertation were addressed with the three studies as examples and their cumulative results presented in this dissertation.

Study 1 presented in Chapter 2, examined the measurement properties of the Social Responsiveness Scale (SRS; Constantino & Gruber, 2005) in an accelerated longitudinal sample of 205 4-year-old preschool children with the complementary approaches of categorical confirmatory factor analysis and Rasch analysis (Duku et al., 2012a). Measurement models based on the literature and other hypothesized measurement models were tested using categorical

confirmatory factor analysis. Rasch analysis was used as a complementary approach that examined unidimensionality and invariance across item and person and over time, as well as convergent validity with other child outcomes.

Study 2 presented in Chapter 3 evaluated alternative measurement models for the autism symptom phenotype based on the Autism Diagnostic Interview-Revised (ADI-R; Rutter, Le Couteur, & Lord, 2003) algorithm items and examined the stability of the most parsimonious and best fitting model across subgroups of interest (Duku et al., 2012b). Data on 3268 individuals aged between 4 and 17 years from the Autism Genome Project (AGP) consortium were used to assess the measurement invariance of the autism symptom phenotype as indexed by the ADI-R across subgroups of participants (younger versus older, verbal versus nonverbal, males versus females). Stability of the autism symptom phenotype was examined using categorical confirmatory factor analysis.

Finally, Study 3, presented in Chapter 4, examined the psychometric properties and measurement structure of the Behaviour Rating Inventory of Executive Functioning-Preschool (BRIEF-P; Gioia, Espy & Isquith, 2003) version using parents' and teachers' reports on 625 children between the ages of 25 and 74 months (Duku & Vaillancourt, 2012). Measurement models were examined using categorical confirmatory factor analysis and invariance of the most parsimonious measurement models across informants were examined. Further analyses included examining the measurement models of the individual clinical scales and the invariance of each of them across informants.

Summary of thesis contributions

With the availability of statistical and methodological approaches focused on measurement invariance there is increasing interest in the evaluation of measurement invariance

in the ECD literature. The type of research conducted in this dissertation is relatively new to this area of research, and human development in general (with the exception cross-cultural research; see Knight & Zerr, 2010). The results of the studies help to confirm the importance of assessing measurement invariance, as well as verifying measurement invariance, in ECD research *prior to* performing group comparisons and interpreting group comparisons (Steinmetz, Schmidt, Tina-Booh, Wieczorek, & Schwartz, 2007). As a recommendation, consideration of measurement invariance should not only be a priori consideration, but should also be an aspect of critical appraisal of studies that aim to make conclusions about the utility of tools.

Chapter 2. Investigating the measurement properties of the Social Responsiveness Scale in preschool children with autism spectrum disorders

J Autism Dev Disord
DOI 10.1007/s10803-012-1627-4

Investigating the Measurement Properties of the Social Responsiveness Scale in Preschool Children with Autism Spectrum Disorders

Eric Duku · Tracy Vaillancourt · Peter Szatmari · Stelios Georgiades · Lonnie Zwaigenbaum · Isabel M. Smith · Susan Bryson · Eric Fombonne · Pat Mirenda · Wendy Roberts · Joanne Volden · Charlotte Waddell · Ann Thompson · Teresa Bennett · the Pathways in ASD Study Team

Abstract The purpose of this study was to examine the measurement properties of the Social Responsiveness Scale in an accelerated longitudinal sample of 4-year-old preschool children with the complementary approaches of categorical confirmatory factor analysis and Rasch analysis. Measurement models based on the literature and other hypothesized measurement models which were tested using categorical confirmatory factor analysis did not fit well and were not unidimensional. Rasch analyses showed that a 30-item subset met criteria of unidimensionality and invariance across item, person, and over time; and this subset exhibited convergent validity with other child outcomes. This subset was shown to have enhanced psychometric properties and could be used in measuring social responsiveness among preschool age children with Autism Spectrum Disorders.

Keywords Social Responsiveness Scale · Autism spectrum disorders · Measurement · Confirmatory factor analysis · Rasch analyses · Structural equation modelling

E. Duku · P. Szatmari · S. Georgiades · A. Thompson · T. Bennett
Offord Centre for Child Studies, McMaster University, Hamilton, ON, Canada

T. Vaillancourt (✉)
Faculty of Education and School of Psychology, University of Ottawa, 145, Jean-Jacques-Lussier Private, Ottawa, ON K1N 6N5, Canada
e-mail: tracy.vaillancourt@uottawa.ca

L. Zwaigenbaum · J. Volden
University of Alberta, Edmonton, AB, Canada

I. M. Smith · S. Bryson
Dalhousie University/IWK Health Centre, Halifax, NS, Canada

E. Fombonne
Montreal Children's Hospital, Montreal, QC, Canada

P. Mirenda
University of British Columbia, Vancouver, BC, Canada

W. Roberts
University of Toronto, Toronto, ON, Canada

C. Waddell
Simon Fraser University, Vancouver, BC, Canada

Introduction

Autism spectrum disorders (ASD) are neurodevelopmental conditions with considerable morbidity and costs to individuals, their families, and society (Charman 2007). ASD affects roughly 1 in 88 preschool aged children (Centers for Disease Control and Prevention 2012). Children are diagnosed with ASD based on impairments in social interaction and communication, as well as a pattern of repetitive or stereotypic behaviour and interests (APA 2000). The diagnosis of ASD usually involves the use of direct observation instruments (e.g., Autism Diagnostic Observation Schedule, ADOS; Lord et al. 2000), parent interviews (e.g., Autism Diagnostic Interview-Revised, ADI-R; Lord et al. 1994), and additional information from independent sources such as teachers, as well as clinical judgment (Newschaffer et al. 2007). Questionnaires such as the Social Responsiveness Scale (SRS; Constantino and Gruber 2005) are frequently used to obtain supplementary information about the child's symptoms.

The SRS is a 65-item quantitative scale that measures the severity of social impairment symptoms related to ASD in individuals between 4 and 18 years of age. In addition to a total severity score, the SRS has five conceptually

Published online: 23 August 2012

derived subscales: Social Awareness, Social Cognition, Social Communication, Social Motivation, and Autistic Mannerisms (Constantino and Gruber 2005). The SRS has been shown to have good sensitivity and specificity and to be informative for differential diagnosis, successfully distinguishing ASD from other childhood psychiatric conditions (Constantino and Gruber 2005). The SRS can be completed in 20 minutes by parents, teachers, or childcare providers who have observed the child's interactions with peers in naturalistic settings. In contrast to other measures of ASD symptoms (e.g., ADOS and ADI-R), the measurement framework for the SRS models autism as a unidimensional phenotype rather than a multidimensional and/or categorical construct (Constantino and Gruber 2005).

Constantino and colleagues have examined the psychometric properties and utility of the SRS across different samples of children varying in age from 4 to 18 years. For example, Constantino et al. (2000) examined the discriminant validity of the SRS in 158 child psychiatric patients, with and without ASD, and a control sample of 287 children randomly selected from a school district. The factor structure of the SRS was also examined using Latent Class Analysis. Results indicated that a one-factor solution, which explained 70 % of the variance, best fit the data. Constantino et al. (2004) re-examined the factor structure using principal component analysis with a clinical sample of 168 children and 259 administrations of the SRS (parent and teacher) and once again showed that a one-factor solution, explaining about 35 % of the variance, best fit the data. Based on these and other analyses, Constantino et al. (2004) concluded that there was no evidence of separate independent subdomains of impairment associated with autism as measured by the SRS in 4- to 18-year-olds. This finding is inconsistent with those reported in previous studies that have described the ASD symptom phenotype as multidimensional, comprising the three domains of social deficits, communication deficits, and fixated interests/repetitive behaviour (Frazier et al. 2008; Georgiades et al. 2007). It should also be noted that the difference in amounts of variance explained could be attributed to the two methodological approaches used—factor analysis (FA) and latent class analysis (LCA).

Although the SRS is commonly used to assess severity of autistic social impairment symptoms in children and youth with ASD, most of the psychometric work has been completed in general population samples. More research has been needed to address the measurement properties of the SRS in children diagnosed with ASD. Specifically, data are needed on the measurement model of the SRS (i.e., uni- vs. multi-dimensionality), and on the longitudinal invariance and psychometric stability of the SRS. It would also be informative to examine the measurement properties of the SRS in younger children, given that ASD is being identified more often in the early preschool years.

Earlier work on the psychometric properties of the SRS used classical test theory approaches such as linear factor analyses, correlations, and item-total correlations (or internal consistency). Although these tests are informative at the scale level, they do not allow the examination of response patterns for individual items. As a complementary approach, Rasch analysis using a latent trait model (with single items treated as indicators) is increasingly being used in health sciences research because it allows researchers to test for response patterns for individual items, for individual-person estimates, and for individual item and person fits and residuals (Hagquist et al. 2009). Rasch's unidimensional measurement model reflects a fundamental feature of measurement; an instrument should work the same way for all individuals (Andrich 1988). Rasch analysis is usually employed in developing and examining measurement instruments, and is useful in analysing the psychometric properties of composite measures that are considered to capture unidimensional constructs, such as the SRS. It is appropriate for Likert-like response items as in the SRS and can be classified as an Item Response Theory (IRT) model. Rasch analysis involves testing whether patterns of responses to items conform to model expectations. To our knowledge, Rasch analysis has not been used to evaluate the measurement properties of instruments assessing ASD symptoms.

The objective of this study was to use multiple methods to investigate the measurement properties of the 4–18-year version of the SRS in a sample of newly diagnosed 4-year-old preschool children with ASD. This objective was achieved by examining: (1) competing measurement models of the SRS in a clinical ASD sample; (2a) an empirical measurement model of the underlying structure of the SRS using Rasch analysis; (2b) the longitudinal invariance of the resulting measurement model of the SRS; and (3) the convergent and discriminant validity of the resulting measurement model with concurrent child outcome measures.

Methods

Participants

Participants were recruited into an on-going longitudinal study of children with ASD (*Pathways in ASD* study) through regional ASD referral centres across Canada (Halifax, Montreal, Hamilton, Edmonton, and the Greater Vancouver/Fraser Valley regions of British Columbia). The study was approved by the Research Ethics Boards at all sites. Families willing to participate provided informed

consent prior to joining the study. Participants included 339 children younger than 60 months (mean age at consent = 39.8 months, SD = 8.9) with a recent diagnosis of ASD. Inclusion criteria for participation in the Pathways Study were as follows: (a) recent (i.e., within 4 months) clinical diagnosis of ASD, confirmed by both the Autism Diagnostic Observation Schedule (ADOS; Lord et al. 2000) and the Autism Diagnostic Interview-Revised (ADI-R; Risi et al. 2006), as well as a diagnosis assigned by a clinician using DSM-IV criteria (APA 2000); and (b) chronological age equal to, or older than, 2 years and equal to, or younger than, 5 years and 0 months. Children were excluded from the study if any of the following conditions were present: (a) cerebral palsy or other neuromotor disorders interfering with study assessments; (b) any known genetic or chromosomal abnormality; or (c) severe visual or hearing impairment. To ensure independence of observations, only one child per family was recruited to the study.

There were both cross-sectional and longitudinal components to this study. For the cross-sectional component, this paper focused on analyses involving an accelerated longitudinal sample (n = 205; 177 boys and 28 girls) of 4-year-olds derived by combining data from baseline (T1; n = 61), 6-month follow-up (T2; n = 75) and one-year follow-up (T3; n = 69), drawn from the whole *Pathways in ASD* cohort. The accelerated longitudinal sample was used so that data could be combined from children of the same age (4 years), from the three "age at time of diagnosis" cohorts at different assessment points (4-year-olds at T1, T2 and T3; see Fig. 1). Each child provided data at only one time point of assessment. For the longitudinal component, the accelerated longitudinal data (T1, T2, and T3) were used as the baseline data along with follow-up data at 6 years (T4). Parents of participants provided ratings for some instruments while other ratings were based on observer reports. Most (93.3 %) parent reports were obtained from mothers, with a mean maternal age at consent of 35.3 years (SD = 5.3).

Measures

Social Responsiveness Scale (SRS)

The SRS provides a picture of a child's atypical social behaviour including social awareness, social information processing, reciprocal communication, social anxiety or avoidance, and autistic preoccupations and traits (Constantino and Gruber 2005). Ratings on the items are provided by the child's caregiver on a scale from 1 (not true) to 4 (almost always true) based on the frequency (not the intensity) of the behaviour. The items vary in degree of abnormality since some inquire about mildly abnormal behaviours whereas others inquire about the severely abnormal. Higher total scores indicate greater severity of social impairment.

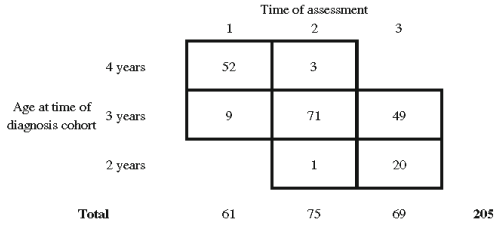


Fig. 1 Accelerated longitudinal design used for children 4 years old at time of assessment

Child Behavior Checklist 1.5–5 (CBCL1.5-5)

The 99-item CBCL is a widely-used norm-referenced instrument that can evaluate a wide range of internalizing and externalizing disorders, based on six subscales (Emotionally reactive, Anxious/depressed, Somatic complaints, Withdrawn, Attention problems, Aggressive behavior; Achenbach and Rescorla 2000). The CBCL is completed by parents or teachers based on observations of the child's behaviour in the previous 2 months. Scale and subscale scores are summed and converted to T-scores. The CBCL has good test–retest and inter-rater reliability for all scales and subscales. The authors also report evidence of discriminative, convergent, and predictive validities (Achenbach and Rescorla 2000).

Repetitive Behavior Scale-Revised (RBS-R)

The RBS-R is a clinical rating scale that measures the presence and severity of a range of restricted, repetitive behaviours that are associated with ASD (Bodfish et al. 2000). It is completed by parents and provides a quantitative, continuous measure of repetitive behaviours. It consists of 43 items distributed across six conceptually derived subscales: Stereotyped behaviour, Self-injurious behaviour, Compulsive behaviour, Routine behaviour, Sameness behaviour, and Restricted behaviour. Mirenda et al. (2010) validated the utility of the RBS-R as a measure of repetitive behaviours in this sample of preschool children with ASD.

Preschool Language Scale: Fourth Edition (PLS-4)

The PLS-4 is a norm-referenced and comprehensive language test for identifying children with a language disorder or delay. It is administered individually to children between birth and age 6 years and 11 months, or to older children who function developmentally within this age

range (Zimmerman et al. 2002). The PLS-4 was used to obtain an index of early syntax and semantic skill in this sample of preschool children with ASD (Volden et al. 2011).

Vineland Adaptive Behavior Scale Second Edition (VABS-II)

The VABS-II was designed to assess functioning from birth to 18 years in the domains of Communication, Daily living skills, Socialization, and Motor skills (Sparrow et al. 2005). Scores from domains and sub-domains permit the comparison of specific profiles of adaptive behaviours in these groups. The VABS-II is administered to parents or caregivers using a semi-structured interview format. Open-ended questions are used to gather detailed information and promote rapport between interviewers and respondents. The VABS-II has been shown to have adequate reliability and validity (Sparrow et al. 2005).

Analyses

The analyses were conducted in three stages. First, the measurement properties of the SRS were examined by considering internal consistencies of the total SRS score and subscales in the sample. Second, the hypothesized unidimensional measurement model and other hypothesized measurement models were evaluated using categorical confirmatory factor analysis (CCFA) in a structural equation modelling (SEM) framework with Mplus version 5.1 (Muthén and Muthén 2008a, b). Tests of goodness-of-fit of the models were evaluated using the criteria described by Hu and Bentler (1999), who recommend conducting several goodness-of-fit tests and reporting their resulting indices. Third, based on the results of the CCFAs, the measurement model of the SRS was examined using Rasch analysis of the 65 items as a complementary approach.

The Rasch model for Likert-like items with ordered categories, called the polytomous Rasch model, is also known as the rating scale model (Andrich 1988). The rating scale model assumes equal intervals between adjacent categories across all items whereas the partial credit model does not impose any restrictions on the intervals (Andrich 1988). The Rasch polytomous model is suitable for analysis of the SRS which has items with ordered Likert-like categories.

The Rasch unidimensional measurement model assumes that the probability that a particular individual will endorse an item is a logistic function of the relative distance between the location of the item and the person location. In other words, the probability that a parent or caregiver will endorse an item is a logistic function of the difference between the child's level or severity of autistic symptoms and the level of severity of autistic symptoms expressed by the item. The response patterns are tested against the expected pattern and a variety of fit statistics are used to determine how well the responses fit the pattern (Hagquist et al. 2009). A good fit of the response pattern means that for the same latent trait, the probability of endorsing a more severe item is higher than the probability of endorsing a less severe item. Rasch's work has been extensively reviewed by other statisticians and methodologists and is well suited to the examination of the measurement properties of instruments such as the SRS (Andrich 2004; Hagquist et al. 2009; Tennant and Conaghan 2007; Wright 1977).

Using the RUMM2020 software (Andrich et al. 2007), the Rasch analysis of the SRS was iterative and included: (a) a test of which polytomous version (rating scale or partial credit model) was appropriate using the Likelihood Ratio Test between models; (b) an overall test of how well the SRS data fit the Rasch model; (c) stepwise deletion of items that showed local dependency based on correlations of residuals over 0.3 after removal of the Rasch model; (d) stepwise deletion of poor-fitting items with extreme item-fit residuals (over $\pm$ 2.5); (e) deletion of items with disordered thresholds (those thresholds between response options of items that do not display an increasing level of the trait); (f) deletion of cases with extreme person-fit residuals (over $\pm$ 2.5), and testing for differential item response patterns or differential item functioning (DIF) across the three data points of the accelerated longitudinal sample; and finally, (g) a test of how well the remaining items and cases fit the Rasch model and the assumptions of invariance and unidimensionality.

Longitudinal invariance was examined with Rasch analysis using the accelerated longitudinal data (T1, T2, and T3) as the baseline and follow-up data at 6 years (T4). The Rasch model was considered to be an adequate fit if the summary and individual χ^2 statistics were non-significant ($p > 0.05$) after adjusting for multiple testing using the Bonferroni correction (Hagquist et al. 2009). Evidence of differential item functioning (DIF) was assessed by analysing the residuals with the three data points and the estimated latent score as covariates (Hagquist and Andrich 2004).

Finally, convergent validity was examined by comparing the resulting total score from the retained items of the SRS to the 65-item total score and to concurrent child outcome measures (the CBCL, VABS-II, RBS-R and PLS-4).

Results

Examining Measurement Models of the SRS

Internal consistency of the 65-item SRS total raw score was good (Cronbach's $\alpha = 0.93$), indicating strong

Table 1 Internal consistencies of the data for the total and subscales raw scores of the SRS in 4-year-olds (n = 205)

Scale/subscale	# Items	Cronbach's alpha
Total	65	0.93
Social awareness	8	0.60
Social cognition	12	0.72
Social communication	22	0.85
Social motivation	11	0.70
Autistic mannerisms	12	0.79

intercorrelations between items and the total raw score. The internal consistencies for the SRS subscales for this sample were also acceptable (above 0.70) except for the Social Awareness subscale, for which the internal consistency was 0.60 (see Table 1).

The goodness-of-fit indices from the CCFA indicated a poor fit for the one-factor (unidimensional) structure comprising all 65 items of the SRS (χ^2 (113) = 431.975, CFI = 0.686, TLI = 0.761, RMSEA = 0.119). Two other tested models, a 5-factor first-order model and a 5-factor second-order model with the subscales as factors, also did not fit well (see Table 2). Single-order unidimensional measurement models were also tested for the SRS subscales. As summarized in Table 2, none of these models met criteria for adequate fit suggested by Hu and Bentler (1999): RMSEA value less than or equal to 0.6, and/or CFI (or TLI) greater than or equal to 0.9, and/or a Chi-square statistic with p value greater than 0.05.

Rasch Analysis: Measurement Model

Since none of the hypothesized measurement models of the SRS provided a good fit to the data, Rasch analysis was used as a complementary approach to examine the measurement properties of the 65-item SRS, including assessing dimensionality, response patterns for individual items, individual person estimates, and individual item and person residuals with fit indices. Results indicated that the partial credit model was appropriate for the data based on the Likelihood Ratio Test ($\chi^2(127) = 434.2$, $p < 0.001$). The person separation index (PSI; a measure of reliability, similar to Cronbach's α) of 0.93 indicated high internal consistency. However, the overall fit of the model evaluated using the latent-trait Chi-square statistic from Table 3 for the 65-item SRS was poor ($\chi^2(130) = 130, p < 0.001$). A test of local dependencies (or intercorrelations) between items based on a residual PCA (after removing the Rasch model) indicated that 10 items had correlations over 0.3 with other items. Using an iterative process, 9 items were initially excluded from the item set because they did not fit the Rasch model (i.e., item-fit residuals greater than ±2.5)

Table 2 Goodness-of-fit statistics for models tested using categorical confirmatory factor analyses with four-year-olds from T1, T2, and T3 (n = 205)

	Chi-sq, *df*, *p* value	CFI	TLI	RMSEA
1-Factor model	431.975, 113, <0.001	0.686	0.761	0.119
5-Factor model	419.516, 113, <0.001	0.698	0.770	0.116
2nd Order 5-factor model	420.940, 113, <0.001	0.697	0.769	0.116
Social awareness	47.515, 15, <0.001	0.767	0.720	0.104
Social cognition	141.071, 28, <0.001	0.654	0.666	0.142
Social communication	264.904, 65, <0.001	0.675	0.823	0.124
Social motivation	162.396, 26, <0.001	0.680	0.693	0.162
Autistic mannerisms	120.732, 32, <0.001	0.833	0.875	0.117

CFI Comparative Fit index, *TLI* Tucker-Lewis Index, *RMSEA* root mean square error of approximation, *SRMR* standardized root mean square residual

or were locally dependent on other items (i.e., residual correlations greater than 0.3). In the second phase, 11 of the remaining 46 items had disordered thresholds and 5 other items did not fit the Rasch model, so an additional 16 items were excluded from the item set, for a total of 35 items. For example, the item "doesn't recognize when others are trying to take advantage of him/her" displayed a fit to the Rasch model whereas another item ("knows when he/she is too close to someone or is invading someone's space") displayed under-discrimination to the Rasch model. Of the 35 items excluded, 5 were excluded from the Social Awareness subscale, 6 from the Social Cognition subscale, 15 from the Social Communication subscale, 5 from the Social Motivation subscale and 4 from the Autistic Mannerisms subscale.

In the final phase of the iterative procedure, 24 children were also excluded from the analyses because their person-fit residuals were over ± 2.5 (i.e., their expected person estimates were 2.5 standardized units away from the observed person estimates). The 30-item subset comprised 8 items which were retained from the Autistic Mannerisms subscale, 7 from the Social Communication subscale, 6 from each of the Social Motivation and Social Cognition subscales, and 3 from the Social Awareness subscale (see "Appendix"). The statistics for the resulting model with 30 items are presented in Table 3. As shown, the item-trait test was non-significant ($\chi^2(60) = 76.008$, $p = 0.08$), indicating that the data fit the Rasch model and that assumptions of invariance (item and person) and unidimensionality held.

Table 3 Summary of test of fit statistic for Rasch analysis of the SRS using selected sample of 205 4-year-olds

	N	Item residual value (SD)	Person residual value (SD)	Item-trait Chi-square	Degrees of freedom	*p* value	Person Separation Index	Power of fit test
65-Item partial credit model	205	0.000 (0.626)	−0.333 (0.577)	248.8	130	<0.001	0.931	Excellent
65-Item rating scale model	205	0.000 (0.763)	−0.381 (0.579)	238.767	130	<0.001	0.931	Excellent
30-Item partial credit model	181	0.000 (0.583)	−0.387 (0.667)	76.008	60	0.080	0.884	Excellent
30-Item rating scale model	181	0.000 (0.636)	−0.391 (0.667)	81.256	60	0.035	0.884	Excellent

Rasch Analysis: Invariance (Across Groups and Over Time)

Examination of DIF for the 3 data points from which 4-year-old participants were drawn (i.e., baseline, 6 and 12 months later) showed no evidence of uniform or non-uniform DIF since no comparisons using analyses of variance were statistically significant. This indicated invariance of the measurement model across the three data points in the accelerated longitudinal sample.

Further examination of DIF using data from the accelerated longitudinal sample (i.e., at age 4) as the baseline and data from time 4 (i.e., 12–36 months later) as follow-up showed that other than uniform DIF for "doesn't recognize when others are trying to take advantage of him/her", there was no evidence of uniform or non-uniform DIF, indicating invariance across time. No comparisons using analyses of variance with Bonferroni correction for multiple comparisons were statistically significant.

Convergent Validity

Convergent validity was investigated by examining the correlations between the 30-item set with concurrent child outcome measures in the study for the children in the accelerated longitudinal sample. The strength of the association between the 30-item subset total score and the 65-item SRS total score ($r = 0.94$, $p < 0.001$) indicated that the 30-item subset total score accounted for approximately 88 % of the variance of the 65-item SRS total score. The 30-item subset total score was positively correlated with the CBCL and RBS-R (r from 0.65 to 0.67, see Table 4) and negatively related with the VABS-II ($r = -0.33$, $p < 0.001$). The PLS-4 had no significant relationship with the 30-item subset total scores ($r = -0.09$, $p = 0.191$).

Table 4 Correlations between total score of the 30-item subset with concurrent child outcome measures for the accelerated longitudinal sample of 4-year-olds

Child outcome measures	30-item subset total raw score
SRS (65-item) Total raw score	0.94**
CBCL internalizing problems: total	0.68**
CBCL externalizing problems: total	0.65**
RBS-R overall mean score	0.67**
VABS-II adaptive behaviour composite standard score	−0.33**
PLS-4 total language standard score	−0.09

** $p < 0.01$ level (2-tailed)

Discussion

This is the first comprehensive study to assess the measurement properties of the SRS in a clinical sample of recently diagnosed 4-year-old preschool children with ASD. It is also the first study to use Rasch modelling to examine the properties of the SRS in an ASD sample. Examination of the measurement properties of the SRS in 4-year-olds is important given that many children are being referred for ASD assessment by this age (Chawarska et al. 2007) and that the SRS may be an informative data source for clinicians. However, poor fit statistics and indices from the hypothesized unidimensional CCFAs showed that the 65-item SRS could not be characterized as unidimensional in our study. The implication of this finding is that one cannot assume measurement equivalence for any measure to be used with ASD children across as wide an age span as the SRS suggests.

Examination of the SRS data using the complementary approach of Rasch analysis also confirmed that the 65-item SRS could not be characterized as unidimensional and that the items did not form a well-fitting measurement model. It is possible that the lack of unidimensionality of the 65-item SRS arose because the covariance between the items could not be explained by a single underlying construct. It is also possible that item properties differed according to some grouping variable or item redundancy or dependency. Another possible reason for the poor fit of the measurement models may be that certain SRS items are less relevant for younger children, e.g., "has good personal hygiene" or "has trouble keeping up with the flow of a normal

conversation". That is, the poor fit could be due to mis-targetting of items to children (i.e., a poor spread of items across the full range of their scores), or that the data could have floor (or ceiling) effects due to poor discrimination of items among younger children (Hagquist et al. 2009).

In the Rasch analysis, 35 items were excluded from the set of 65 because of local dependency, lack of fit, and disordered thresholds, all of which could contribute to the lack of adequate fit of the hypothesized unidimensional factor structure. Local dependency means that some items were highly correlated with other items and led to a lack of fit based on tests of residuals (person and item). There were disordered thresholds in 11 items, indicating that the response scale was not functioning as it should and that the meaning of the responses for those items was unclear.

Using Rasch analysis, we also showed that the 65-item SRS could be reduced to a 30-item subset with good internal consistency using data from 4-year-olds. Examination of DIF for the 3 data points using the 30-item subset showed no evidence of uniform or non-uniform DIF. The 30-item subset was shown to be unidimensional and a well-fitting measurement model for the 30-item set, explaining about 88 % of the variance of the 65-item SRS. Post hoc examination of the 30-item subset using the full sample showed that the data were a good fit to the Rasch model and the internal consistency based on the PSI was also high. Examination of concurrent validity of the 30-item subset total score with child outcome measures showed that the total score was positively associated with CBCL subscales and RBS-R domains, indicating that severity of autistic social impairment was associated with severity of internalizing/externalizing behaviour and repetitive behaviours. Similar to the findings of Constantino and Gruber (2005), there was a negative association of the SRS with adaptive functioning (VABS-II Adaptive Behavior Composite score) but no statistically significant relationship with language skills (PLS-4 Total score).

The 30-item subset may prove useful for research in preschool samples, as it is easier to implement and yields a single-dimension construct (as proposed by Constantino and Gruber 2005). We were also able to show the utility of Rasch analysis as a complementary approach in examining the measurement properties of the SRS and in the refinement of the SRS. Indeed, although the 65-item SRS did not perform well statistically, possibly because of the age of the children, the 30-item subset appeared to represent "markers" of autistic social impairment that functioned well across age groups, even in the narrow follow-up interval ranging from 12 to 36 months. The 30-item subset also meets the assumptions underlying the Rasch model and therefore may have potential for use in evaluating the severity of autistic social impairment as a single dimension in other clinical samples of preschool children. These results suggest the need for the findings be replicated in larger independent samples with wider age ranges. Independent studies should also evaluate the psychometric properties and clinical utility of the 30-item subset.

One limitation of this study is that the SRS was designed for use in 4- to 18-year-olds, yet our data came from preschoolers at the lower end of the age range (4 years). Another limitation was that fewer than 100 4-year-olds were available for analyses at any single time point. We were able to test 205 4-year-olds in this study by creating an accelerated longitudinal dataset by combining data from three time points. As a consequence, age (or cohort effects) could influence our results.

In conclusion, our findings suggest that the structure of the 65-item SRS cannot be described as unidimensional in this sample of 4-year-olds with ASD. Moreover, a substantial number of SRS items functioned poorly in not discriminating well among preschool children with ASD, at least in this sample. The complementary Rasch analysis showed that the 65-item SRS could be reduced to a 30-item subset with little loss in explanatory power, with the relationships between the 30-item subset and other outcome measures being similar to those found with the 65-item SRS. Use of this subset of SRS items is therefore recommended in measuring the severity of social impairment among preschool age children with ASD.

Acknowledgments This study was supported by the Canadian Institutes of Health Research, Autism Speaks, the Government of British Columbia, Alberta Innovates—Health Solutions, and the Sinneave Family Foundation. The authors thank all the families who participated in the *Pathways in ASD* study. The authors also acknowledge the members of the *Pathways in ASD Study Team*. These members had equal contribution to the study and are listed here alphabetically: Liliana Abruzzese, Megan Alexander, Susan Bauld, Ainsley Boudreau, Colin Andrew Campbell, Mike Chalupka, Lorna Colli, Melanie Couture, Bev DaSilva, Vikram Dua, Miriam Elfert, Lara El-Khatib, Lindsay Fleming, Kristin Fossum, Nancy Garon, Shareen Holly, Stephanie Jull, Karen Kalynchuk, Kathryne MacLeod, Preetinder Narang, Julianne Noseworthy, Irene O'Connor, Kaori Ohashi, Jennifer Endre Olson, Sarah Peacock, Teri Phillips, Sara Quirke, Katie Rinald, Jennifer Saracino, Cathryn Schroeder, Cody Shepherd, Rebecca Simon, Mandy Steiman, Richard Stock, Benjamin Taylor, Lee Tidmarsh, Larry Tuff, Kathryn Vaillancourt, Stephen Wellington, Isabelle Yun, and Li Hong Zhong.

Appendix

See Table 5.

Table 5 30-item subset of SRS and original subscales

Item	Subscale	Description
SRS1	SMot	Fidgety in social situations than when alone
SRS4	AMan	Under stress, shows rigid or inflexible patterns of behavior
SRS5	SCog	Doesn't recognize when others take advantage
SRS6	SMot	Would rather be alone
SRS9	SMot	Clings to adults
SRS14	AMan	Not well coordinated
SRS15	SCog	Understands meaning of people's tone and facial expressions
SRS16	SCom	Avoids eye contact
SRS19	SCom	Gets frustrated trying to get ideas across in conversations
SRS20	AMan	Shows unusual sensory interests
SRS22	SCom	Plays appropriately
SRS23	SMot	Does not join group activities
SRS24	AMan	Difficulty with changes in routine
SRS25	SAw	Doesn't seem to mind being out of step
SRS29	AMan	Regarded by other children as odd
SRS30	SCog	Becomes upset in a situation with lots of things going on
SRS31	AMan	Can't get mind off something
SRS33	SCom	Socially awkward
SRS36	SCom	Difficulty relating to adults
SRS42	SCog	Overly sensitive to sounds
SRS44	SCog	Doesn't understand how events relate to one another
SRS46	SCom	Serious facial expressions
SRS49	AMan	Does extremely well at a few tasks,
SRS52	SAw	Knows when talking too loud
SRS56	SAw	Walks between two people talking
SRS57	SCom	Teased a lot
SRS58	SCog	Concentrates on parts rather than seeing the whole picture
SRS63	AMan	Touches in unusual way
SRS64	SMot	Too tense in social settings
SRS65	SMot	Stares or off into space

AMan autistic mannerisms, *SMot* social motivation, *SCog* social cognition, *SCom* social communication, *SAw* social awareness

References

Achenbach, T. M., & Rescorla, L. A. (2000). *Manual for the ASEBA preschool forms & profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families.

American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders DSM-IV-TR* (4th ed.; text revision). Washington, DC: American Psychiatric Association.

Andrich, D. (1988). *Rasch models for measurement: Sage University Paper Series on Quantitative Measurement in the Social Sciences*. Newberry Park, CA: Sage.

Andrich, D. (2004). Controversy and the Rasch model. *Medical Care, 42*(1 Suppl), i7–i16.

Andrich, D., Sheridan, B., & Luo, G. (2007). *RUMM2020: Rasch unidimensional measurement models software, version 4.1*. Perth, Western Australia: RUMM Laboratory.

Bodfish, J. W., Symons, F. J., Parker, D. E., & Lewis, M. H. (2000). Varieties in repetitive behavior in autism. *Journal of Autism and Developmental Disorders, 30*, 237–243.

Centers for Disease Control and Prevention. (2012). Prevalence of autism spectrum disorders—Autism and developmental disabilities monitoring network, 14 sites, 2008. *Morbidity & Mortality Weekly Report Surveillance Summaries, 61*(3), 1–19.

Charman, T. (2007). Autism and its impact on child development. In R. E.Tremblay, R. G. Barr, & R. D. V. Peters (Eds.). *Encyclopedia on early childhood development* [online] (pp. 1–5). Montreal, Quebec: Centre of Excellence for Early Childhood Development. Retrieved November 8 2011 from http://www.child-encyclopedia.com/documents/CharmanANGxp_rev.pdf.

Chawarska, K., Paul, R., Klin, A., Hannigen, S., Dichtel, L. E., & Volkmar, F. (2007). Parental recognition of developmental problems in toddlers with autism spectrum disorders. *Journal of Autism and Developmental Disorders, 37*, 62–72.

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

Constantino, J. N., Gruber, C. P., Davis, S., Hays, S., Passante, N., & Przybeck, T. (2004). The factor structure of autistic traits. *Journal of Child Psychology and Psychiatry, 45*, 719–726.

Constantino, J. N., Przybeck, T., Friesen, D., & Todd, R. D. (2000). Reciprocal social behavior in children with and without pervasive developmental disorders. *Journal of Developmental and Behavioral Pediatrics, 21*(1), 2–11.

Frazier, T. W., Youngstrom, E. A., Kab, C. S., Sinclair, L., & Rezai, A. (2008). Exploratory and confirmatory factor analysis of the Autism Diagnostic Interview-Revised. *Journal of Autism and Developmental Disorders, 38*, 474–480.

Georgiades, S., Szatmari, P., Zwaigenbaum, L., Duku, E., Bryson, S., Roberts, W., et al. (2007). Structure of the autism symptom phenotype: A proposed multidimensional model. *Journal of the American Academy of Child and Adolescent Psychiatry, 46*, 188–196.

Hagquist, C., & Andrich, D. (2004). Is the sense of coherence-instrument applicable on adolescents? A latent trait analyses using Rasch-modelling. *Personality and Individual Differences, 36*, 955–968.

Hagquist, C., Bruce, M., & Gustavsson, J. P. (2009). Using the Rasch model in nursing research: An introduction and illustrative example. *International Journal of Nursing Studies, 46*, 380–393.

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

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B., DiLavore, P. C., et al. (2000). The Autism Diagnostic Observation Schedule-Generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders, 30*, 205–233.

Lord, C., Rutter, M., & Le Couteur, A. (1994). Autism Diagnostic Interview-Revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders, 24*, 569–685.

Mirenda, P., Smith, I., Vaillancourt, T., Duku, E., Georgiades, S., Szatmari, P., et al. (2010). Validating the Repetitive Behavior Scale-Revised in young children with autism spectrum disorder. *Journal of Autism and Developmental Disorders, 40*, 1521–1530.

Muthén, L. K., & Muthén, B. (2008a). *Mplus 5.1 for Windows*. Los Angeles, CA: Author.

Muthén, L. K., & Muthén, B. (2008b). *Mplus user's guide*. Los Angeles, CA: Author.

Newschaffer, C. J., Croen, L. A., Daniels, J., Giarelli, E., Grether, J. K., Levy, S. E., et al. (2007). The epidemiology of autism spectrum disorders. *Annual Review of Public Health, 28*, 235–258.

Risi, S., Lord, C., Gotham, K., Corsello, C., Chrysler, C., Szatmari, P., et al. (2006). Combining information from multiple sources in the diagnosis of autism spectrum disorders. *Journal of the American Academy of Child and Adolescent Psychiatry, 45*(9), 1094–1103.

Sparrow, S. S., Cicchetti, D. V., & Balla, D. A. (2005). *Vineland II: A revision of the Vineland adaptive behavior scales: I. survey/ caregiver form*. Circle Pines: American Guidance Service.

Tennant, A., & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis and Rheumatism, 57*(8), 1358–1362.

Volden, J., Smith, I. M., Szatmari, P., Bryson, S., Fombonne, E., Mirenda, P., et al. (2011). Using the preschool language scale, fourth edition to characterize language in preschoolers with ASD. *American Journal of Speech-Language Pathology, 20*, 200–208.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*(2), 97–116.

Zimmerman, I., Streiner, R., & Pond, R. (2002). *Preschool language scale* (4th ed.). San Antonio, TX: Psychological Corp.

Chapter 3. Measurement equivalence of the autism symptom phenotype in children and youth

Measurement equivalence of the autism symptom phenotype in children and youth

Eric Duku[1], Peter Szatmari[1], Tracy Vaillancourt[2], Stelios Georgiades[1], Ann Thompson[1], Xiao-Qing Liu[3], Andrew D. Paterson[4, 5], Terry Bennett[1]

[1]Offord Centre for Child Studies & Department of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton, ON Canada; [2]Faculty of Education, University of Ottawa, Ottawa, ON, Canada; [3]Department of Obstetrics, Gynaecology, and Reproductive Sciences, University of Manitoba, Winnipeg, MB, Canada; [4]Genetics and Genome Biology, The Hospital for Sick Children, Toronto, ON Canada; [5]Dalla Lana School of Public Health, University of Toronto, Toronto, ON Canada;

Abstract

Background: The Autism Diagnostic Interview-Revised (ADI-R) is a gold standard assessment of Autism Spectrum Disorder (ASD) symptoms and behaviours. A key underlying assumption of studies using the ADI-R is that it measures the same phenotypic constructs across different populations (i.e. males and females, younger and older, verbal and nonverbal). The objectives of this study were to evaluate alternative measurement models for the autism symptom phenotype based on the ADI-R algorithm items and to examine the stability of the most parsimonious and best fitting model across subgroups of interest. **Methods**: Data came from the Autism Genome Project consortium and consisted of 3628 children aged 4 to 18 years (84.2% boys and 75% verbal). Twenty-eight algorithm items applicable to both verbal and non-verbal participants were used in the analysis. Stability of the autism phenotype was examined using categorical confirmatory factor analysis. **Results**: The autism symptom phenotype was best indexed by the first-order, 6-factor measurement model proposed by Liu et al. (2011). This model was well fitting and stable across subgroups of participants (median age, verbal ability and sex). An alternative second-order model resembling the proposed DSM-5 2-factor structure of the phenotype also showed good overall fit but not for all the subgroups. **Conclusions**: The autism symptom phenotype is adequately characterized by a 6-factor measurement model; this model appears to be stable across subgroups of children and youth with ASD. The 2-factor model provides equally good fit for the sample as a whole but comparison of these two dimensions between subgroups that differ in age, sex and verbal ability is made difficult by lack of measurement equivalence. The derivation of a 6-factor model that presents a good fit to the data and is equivalent across subgroups of interest allows for the use of the derived first-order quantitative traits in future studies of etiology, outcome and treatment.

Measurement equivalence of the autism symptom phenotype in children and youth

There has been considerable debate concerning the underlying structure of the autism phenotype since the seminal article by Kanner (1943) who described the fundamental structure of autism as being composed of "aloneness" and "insistence on sameness". The key question for many researchers and clinicians is the extent to which the clinical manifestation(s) of Autism Spectrum Disorder (ASD) represent multiple underlying domains or a single "continuum". The Diagnostic and Statistical Manual - 4th edition (DSM-IV) conceptualized ASD as a "triad of symptoms" consisting of impairments in social reciprocity, in verbal and nonverbal communication, and a pattern of repetitive stereotyped behaviours (APA, 2000). More recently, the proposed DSM-5 conceptualization of ASD has been influenced by empirical studies suggesting that the symptom structure is best represented by two dimensions: social communication deficits and fixated interests and repetitive behaviours (APA, 2010). However, this does not rule out the possibility that there might be other relevant and informative factors or dimensions subsumed under these two overarching domains.

An empirical approach to exploring the underlying structure of the ASD phenotype is based on the use of statistical techniques such as structural equation modeling (e.g., factor analysis, latent class analysis, cluster analysis). There have been many published factor analytic studies using the ADI-R, an instrument viewed as a gold standard for the assessment of ASD symptoms and behaviours (Bölte & Poustka, 2001; Frazier, Youngstrom, Kubu, Sinclair, & Rezai, 2008; Georgiades et al, 2007; Lecavalier et al, 2006; Liu et al, 2011; Prior et al., 1998; Snow, Lecavalier & Houts, 2009; Szatmari et al, 2002; Tadevosyan-Leyfer et al, 2003; van Lang et al, 2006). The number of factors identified in these studies varies anywhere from one to six. This variability in the number of factors might be due to issues related to the composition of the

sample used, the selection of ADI indicators (i.e., total number, domains versus items, etc.), and the criteria by which different models are evaluated and chosen.

To date, three studies have examined the factor structure of the ASD phenotype using only the algorithm items of the ADI-R. Lecavalier et al. (2006) examined the phenotypic structure of these items in a sample of 226 verbal children from the Autism Genetic Resource Exchange (AGRE) database using exploratory factor analysis (EFA) and all 34 algorithm items. Results from this study produced a three-factor solution similar to the original algorithm domains, although all non-verbal communication items loaded on the social factor. Snow, Lecavalier and Houts (2009) examined the ADI-R algorithm items separately in 1,329 verbal and 532 non-verbal children age 4 to 18 years (34 items and 28 items respectively) from AGRE. A two factor model (social/communication items and restricted/repetitive behaviours) explained the ASD symptomatology independent of the verbal ability of the participants. More recently, Liu et al. (2011) investigated the ASD symptom phenotype in a sample (N=2040) from the Autism Genome Project (AGP) using factor analysis of 28 algorithm items of the ADI-R that apply to all children regardless of their verbal ability. Liu et al. identified six factors which they labeled joint attention, social interaction and communication, non-verbal communication, repetitive sensory-motor behaviour, peer interaction, and compulsion/ restricted interests.

Participants in the studies discussed above varied with respect to characteristics such as chronological and developmental age, sex, language ability, and adaptive functioning. These are important phenotypic characteristics that appear to influence or are associated with the heterogeneity of ASD (Hus, Pickles, Cook, Risi Jr, & Lord, 2007; Szatmari et al, 2012). To date, ASD researchers have focused on the comparison of scores/scales across different subgroups. Therefore it is important to simultaneously examine the stability (or "measurement equivalence")

of the structure of the ASD symptom phenotype across subgroups; such an investigation would test the key underlying assumption of studies using the ADI-R which is that the instrument measures the same phenotypic construct *in all subgroups*.

Measurement Equivalence (or invariance) refers to an instrument's ability to measure the same construct in the same way across populations or groups (Millsap & Kwok, 2004). In ASD research, measurement equivalence issues have been investigated by comparing means and variances across different subgroups. For example, boys and girls have been compared on autistic symptoms including social reciprocity, and repetitive stereotyped behaviour (Hoffman, 2009; Szatmari et al, 2012). However, comparing means and variances, although useful for answering specific questions, is not the same as comparing the underlying measurement structure (i.e., testing measurement equivalence) across sub-groups (Vandenberg & Lance, 2000).

A central principle of measurement equivalence is that the scores or ratings across groups are on the same scale and that the empirical relationships between test items and traits of interest must remain stable across groups (Reise, Widaman & Pugh, 1993). In other words, the measure relates to the construct of interest -- in this case the ASD phenotype -- in a stable way, even across quite different groups. In this context, an investigator might be interested in whether the number of factors and the items that make up a factor are similar across important subgroups. Restricting the comparison to testing differences in means (the traditional approach used in ASD research to date) assumes an equivalent underlying measurement model or phenotypic structure across sub-groups which may be an erroneous assumption. It is therefore essential to investigate whether subgroups differ on *both* their means *and* their measurement structure (Gregorich, 2006).

Whether variability in the clinical presentation of ASD is an indication of the variation in number of underlying dimensions has been a subject of some debate (Kamp-Becker, Ghahreman, Smidt, & Remschmidt, 2009). Hence there is the need to establish the dimensional structure of the autism phenotype and to examine its stability across subgroups of ASD children. To our knowledge, none of the studies investigating the derived measurement structure of ASD using either the full set of items or the algorithm items of the ADI-R have investigated the equivalence of the phenotype across subgroups of interest. Ensuring measurement equivalence (or stability) of the factor structure across subgroups of ASD such as verbal ability, sex and age is important so that potential differences (or similarities) between the groups can be interpreted reliably (Muthén, 1989; Vandenberg & Lance, 2000). Examination of stability across age (i.e., over development) is particularly important since ASD is often identified in the early preschool years (Zwaigenbaum, 2010) and (in most cases) continues throughout the lifespan. Measurement equivalence can be tested by fitting increasingly restrictive multi-group models and then examining modification indices, fit statistics and parameters to determine the sources of possible lack of fit.

The main objectives of this study were to evaluate competing measurement models for the autism symptom phenotype based on the ADI-R algorithm items and to examine the stability of the most parsimonious and best fitting model of this phenotype across subgroups of individuals with ASD that differ in terms of verbal ability, sex, and age. Identification of good fitting measurement models that are equivalent across subgroups of interest might inform further research that involves comparison of groups that might vary by age, sex or verbal/non-verbal status.

Methods

Participants

Data for this study came from the Autism Genome Project (AGP), a collaborative research consortium, of scientists from Europe and North America studying genetic mechanisms underlying autism susceptibility. Informed parental consent was obtained for all participants in the study, and institutional review boards approved the research procedures. The sample was selected as follows: to ensure independence, one random child from a family with autism or ASD based on the ADI-R diagnostic criteria (Risi et al., 2006) aged between 4 and 18 was selected from each of the 4644 families in the combined dataset from AGP Phases I and II. The selected sample had to have data on (a) all 28 algorithm items that were aimed at both verbal and nonverbal children as well as (b) the subgroup variables: verbal ability, sex and age.

The final sample selected consisted of 3628 children from 3628 families, aged 4 to 18 years from each of the sites that contributed data to the AGP consortium database. The sample comprised 3054 boys (84.2%) with a median age of 88 months (mean = 99 months, SD = 42 months, minimum = 48 months, maximum=226 months, 25th percentile = 66 months, and 75th percentile = 123 months). Seventy-five percent of the selected sample was classified as verbal. There was an overlap of 35% (n=1277) between the sample used for the initial analysis in the current study and the one used by Liu et al. (2011). The measurement model was tested independently in the unique 65% (n=2351) sample with good fit ($\chi^2(208) = 1462.116$; CFI=0.939, TLI=0.977, RMSEA=0.051) as a prerequisite to the analysis and the two samples were combined for use in this study.

Measures

*Autism Diagnostic Interview-Revised (ADI-R; Rutter, Le Couteur, & Lord, 2003)***.** The ADI-R is a 93-item[1] standardized, semi-structured clinical interview for caregivers of children and adults and is used as a diagnostic instrument for assessing ASD in children and adults and provides a diagnostic algorithm for autism as described in both the ICD-10 and DSM-IV. The instrument is administered in an interview and focuses on behaviour in three main areas: (1) qualities of reciprocal social interaction; (2) communication and language; and (3) restricted and repetitive, stereotyped interests and behaviours. The ADI-R is appropriate for children and adults with mental ages from about 18 months and above. A rating score for each question is determined based on the evaluation of the caregiver's response. A total score is then calculated for each of the content areas and an autism diagnosis is indicated when scores in all three behavioural areas exceed the specified minimum cutoff scores (Rutter, Le Couteur, & Lord, 2003). Published psychometric results show both reliability and validity of the ADI-R using inter-rater reliability, test-retest reliability, and internal validity tests. Inter-rater reliability and internal consistency were good across all behavioural areas in the ADI-R (Lord et al., 1994). The ADI-R has been found to have adequate reliability across time (Lord et al., 1994). A decision was made to exclude items (mostly on verbal communication) that were not applicable to non-verbal children, so as not to exclude such children. The Friendships and Inappropriate Facial Expressions items were also excluded from this analysis since they are not asked of all children in the age span we covered. Scores of 3 were not recoded to 2 as is done for the algorithm totals as this would have reduced the variability in scores.

Analysis

[1] 34 of the 94 items are used in algorithm for diagnosis of autistic disorder based on the DSM-IV diagnostic criteria

Several measurement models were examined. Specifically, we examined: (1) a one-dimensional measurement model of the selected 28 items (2) the six-factor model (see Figure 1) proposed by Liu et al. (2011) allowing for cross-loading items, (3) other competing models – (3a) two-factor model based on the findings of Snow et al. (2008); (3b) a three-factor model based on original DSM-IV domains and (3c) a second-order measurement model based on the model in (2) where the six factors are explained by two higher order factors, in accordance with the proposed DSM-5 definition (see Appendix 1). Since the items of the ADI-R are categorical/ordinal, the data were examined using categorical confirmatory factor analysis (CCFA) in Mplus (version 5.1) which has the facility to handle this type of data by using a robust weighted least squares estimator (Muthén, 1984; Muthén, du Toit, & Spisic, 1997).

Model suitability and fit using the CCFA were assessed based on multiple indices as suggested by Hu and Bentler (1999). For this study we used the Tucker Lewis Index (TLI), Comparative Fit Index (CFI) and the Root Mean Square Error of Approximation (RMSEA) to evaluate model fit. The TLI and CFI measure the proportionate improvement in fit as one progresses from the baseline model to the target model, per degrees of freedom. Hu and Bentler (1999) recommend a cut-off value on either the TLI or CFI of .95. However, a cut-off of 0.92 on the TLI or CFI has also been shown to be acceptable (Byrne, 2010). Hu and Bentler also recommend a cut-off value of .06 for the RMSEA. This is an estimate of how well the fitted model approximates the population covariance matrix per degrees of freedom. All three indices are minimally affected by sample size (Hu & Bentler, 1999).

Next, the fit of the preferred measurement models selected based on parsimony, were evaluated in each of the subgroups of the sample. Subgroups were created by splitting the sample on verbal ability based on the response to the ADI-R item 30 (verbal versus nonverbal),

sex (male versus female), and based on the median age in months (greater than or equal to 88 months versus less than 88 months) to ensure samples of equal size. Based on the results of the fit of the preferred measurement models, the equivalence of the measurement model(s) was examined across subgroups within a structural equation modeling framework using multiple group categorical confirmatory factor analyses (MGCCFA).

MGCCFA is an extension of multiple group confirmatory factor analyses (MGCFA), which is the most widely used method to test for measurement equivalence for items with continuous responses. Based on the seminal work by Karl Joreskog (1971), MGCFA permits testing for measurement equivalence by setting cross-group constraints and comparing more restricted with less restricted models (Baumgartner & Steenkamp, 1998; Beckstead, Yang & Lengacher, 2008; Byrne, Shavelson, & Muthén., 1989). For categorical outcomes, measurement equivalence models constrain thresholds and factor loadings in tandem because the item probability curve is influenced by both parameters (Muthén, 2008).

Muthén and Asparhouhov (2002) suggest performing MGCCFA in several steps. First, thresholds and factor loadings are unconstrained across groups, while the scale factors are fixed at one and factor means fixed at zero in all groups (unconstrained model). Next, thresholds and factor loadings are constrained to be equal across the groups with the scale factors fixed at one for one group and freed in the others and factor means fixed at 0 in one group and freed in the other groups (constrained model). A difference of chi-square test is computed where a significant chi-square means the hypothesis that the constrained parameters are invariant across groups (i.e., the equality constraints are violated) is rejected. Typically, a non-significant chi-square test supports the equivalence of the measures, although it has been shown that with large samples, this is difficult to achieve (Byrne, 2010). Therefore, we used the criterion that change

in both the CFI and TLI was less than 0.01 when comparing the baseline model with the models where parameters were not constrained, suggesting evidence of equivalence (Byrne, 2010). If evidence of lack of measurement equivalence was found, then a test for partial measurement equivalence was performed by further relaxing some of the constraints on parameters in the restricted model (Muthén & Muthén, 2008). We relied on the use of the concept of partial equivalence (Byrne et al., 1989) to assess the equivalence across subgroups.

Byrne et al. (1989) have argued that by implementing *partial measurement invariance*, multi-group analyses can still continue given that the recommended conditions are met. They showed that even when a subset of parameters is constrained to be invariant across groups, cross-group comparisons can be made in the absence of full invariance. Partial invariance can be assessed in 2 ways (1) when parameters are invariant across some but not all groups and (2) when some but not all parameters are invariant across all groups (Vandenberg & Lance, 2000). Thus this allows for some items of the construct to show non-equivalence so long as the latent construct under investigation can be defined equally across the two groups. Constraints were thus relaxed on an item indicating a possible difference in the way the item pertains to children in the subgroups.

Results

Categorical Confirmatory Factor Analysis

From the results of the CCFA with the competing models presented in Table 1, the fit statistics for the first-order six-factor model was better than the fit statistics for the first order 1-, 2- and 3-factor models. The second-order measurement model (6 factors subsumed by 2) based on the DSM-5 proposal had similar fit to the six factor model, and was also better than the first-order 1-, 2-, and 3-factor models. Using the criteria of parsimony and fit indices, the first-order

measurement model (based on the model by Liu et al., 2011) with six factors for the autism phenotype and the second-order DSM-5 measurement model were selected for use in the subsequent analyses (Liu et al.: $\chi^2(220) = 2286.084$; CFI=0.937, TLI=0.977, RMSEA=0.051 and DSM-5: $\chi2(224)=2463.441$; CFI=0.932, TLI=0.975, RMSEA=0.052). Fit statistics and parsimony were the only criteria used since the models were not nested and no statistical test could be used.

Next using CCFA, the selected measurement models for the autism phenotype were examined for fit in each of the subgroups and the results are presented in Table 2. This is a requirement before any further analyses could be performed. The fit of the first-order 6-factor model (Liu et al, 2011) was adequate and met our criteria for each of the subgroups. This suggested that there was evidence that the selected measurement model fit each of the subgroups and we could proceed to testing the equivalence of this model across the subgroups. However, the fit of the second-order selected DSM-5 model did not meet our criteria for all the subgroups (see Table 2). The model did not converge for the test for girls and there was a further challenge with correlations between the latent variables for the age greater 88 month group. Hence the Liu et al. 6-factor model was used for the rest of the analyses.

Measurement Equivalence

Using the Muthén and Asparhouhov (2002) approach to testing measurement equivalence for categorical indicators across groups based on verbal ability, the overall fit of the baseline model with no constraints on factor loadings and thresholds (or unconstrained model) was adequate for the selected measurement model of the autism phenotype (see Table 3). The overall fit of the second model in which the factor loadings and thresholds were constrained to be equal (or the constrained model) was also adequate and is also shown in Table 3. The differences in

fit indices for the test of measurement equivalence of the model were not within the acceptable limits of the criteria suggesting lack of evidence of equality of factor loadings and thresholds (DIFFTEST: $\chi^2(43) = 529.584$; ΔCFI=0.014, ΔTLI=0.007, ΔRMSEA=-0.007). Next, based on evidence from the modification indices, partial equivalence was achieved by relaxing the equality constraints for threshold and factor loadings for one of the indicators of repetitive sensory-motor behaviour ("use of other's body to communicate"*),* resulting in a better fitting model. The results of the test in difference of models suggested evidence of partial equivalence and equality of factor loading and thresholds across verbal ability subgroups (DIFFTEST: $\chi^2(40) = 390.452$; ΔCFI=0.007, ΔTLI=0.004, ΔRMSEA=-0.004).

Results of the examination of the equivalence of the measurement model for the autism phenotype across sex are presented in Table 3. Here again, based on the computed fit indices, the unconstrained and constrained models tested were well fitting. The test of difference in models showed that the hypothesis of equality of factor loadings and thresholds across sex could not be rejected (DIFFTEST: $\chi^2(47) = 156.281$; ΔCFI=0.004, ΔTLI=0.001, ΔRMSEA=-0.002). Thus we could not reject the hypothesis that the measurement model of the ASD symptom phenotype was similar across boys and girls.

Finally, from Table 3, results of the assessment of equivalence across age groups show that both the constrained and unconstrained models have fit statistics that met our fit criteria. The test of difference in models shows that we cannot reject the hypothesis of similarity of models and consequently equality of factor loading and thresholds. (DIFFTEST: $\chi^2(46) = 296.559$; ΔCFI=0.000, ΔTLI=-0.001, ΔRMSEA=-0.001). This result indicates the similarity of the measurement model of the ASD symptom phenotype across the two age groups.

Comparison of subgroups

Based on the evidence of measurement equivalence, descriptive statistics for the factors of the selected measurement model for the autism phenotype are presented in Table 4. Mean factor scores were statistically significantly higher for non-verbal children than verbal children (p<0.001) except for compulsion/restricted interests where the reverse was true. Effect sizes of the differences were medium to large and are also presented in Table 4. Mean factor scores for joint attention, social interaction and communication, non-verbal communication, and peer interaction were significantly higher for children older than 88 months of age compared to children 88 months or younger. Significantly lower factor means were observed for children older than 88 months for repetitive sensory-motor behaviour compared to children 88 months or age or younger. Finally, boys had a significantly lower mean score on non-verbal communication than girls and a higher mean score on compulsion/restricted interests compared to girls. It should be noted that verbal participants were on the average 22 months older than non-verbal participants. A caveat also needs to be added-- these findings could differ if the other grouping variables were analyzed.

Discussion

This study is, as far as we are aware, the first to examine the equivalence of the measurement model of the autism symptom phenotype across subgroups using algorithm items from the ADI-R. This was achieved using categorical symptom indicators within the structural equation modeling framework in a large combined dataset of children and youth with ASD. Twenty-eight algorithm items applicable to both verbal and non-verbal participants were used as they are relevant to the diagnosis of ASD across a wide age range and developmental ability.

Results showed that compared to multiple, alternative measurement models of the autism phenotype previously identified in the literature, the best fitting and most parsimonious model is

the 6-factor model derived by Liu et al. (2011). The results from this study also illustrate the importance of assessing measurement equivalence prior to making comparisons between ASD subgroups. Measurement equivalence of the ADI-R should be a prerequisite for making any meaningful comparisons across groups since the latent constructs we use to index symptoms cannot be directly measured (Borsboom, 2006). Ideally, the differences in scores should reflect true differences in the latent construct. Stability and therefore accuracy of scientific inferences are threatened when there are measurement problems or biases. Furthermore, the credibility of inferences regarding the developmental process or developmental differences in ASD is affected by the absence of established measurement equivalence (Knight & Zerr, 2010).

Study findings support the continuity of the autism symptom phenotype across verbal ability, sex, and age. The results also suggest that the autism phenotype in this sample is best characterized by a six factor measurement model which is stable across subgroups of ASD children. Evidence of a stable structure of the autism phenotype ensures we can be confident that the instrument is measuring the same construct across subgroups of the ASD population in genetic analysis, studies of outcome or response to treatment. Stability of the autism phenotype also enables us to make more meaningful interpretations of individual progress.

It was also true that the 2-factor model provided a good fit to the data for the ASD group as a whole. This supports the conclusion of the DSM 5 Working Group that the ASD phenotype is primarily a dyad of dimensions. The challenge comes when one wants to sub-group children with ASD into more homogeneous clusters. One cannot assume measurement equivalence of the dimensions in those clusters if they also differ by age, sex, or verbal ability. So while the 2-factor model may be good for diagnosis and classification, for purposes that involve comparison of sub-groups, the 6-factor model may be preferred as it demonstrates stability across age, sex, and

verbal ability. This conclusion points to the challenges of establishing invariance of a measurement model for the autism phenotype for both research and for clinical and diagnostic purposes. To be able to establish a dual purpose measurement model using 28 items out of the battery of items used for clinical diagnosis is daunting at the very least, if not impossible.

There is clearly a need for replicating (or extending) our findings from the analysis of ADI-R to other measures of the ASD phenotype such as the Social Responsiveness Scale (Constantino & Gruber, 2005) and the Autism Diagnostic Observation Schedule (Lord et al, 2000). If our findings extend to other instruments that vary by method and item pool, then the potential incompatibility of asking a single tool to cover both clinical and research purposes becomes a significant challenge for the field.

Limitations

This study had some limitations that are worth noting. First, there are differences in the inclusion/exclusion criteria (e.g., age at assessment and severity of symptoms) used by the different sites of the AGP consortium that may have influences our results. Second, this is a research sample with a high median age at administration of the ADI-R. The AGP data were collected for purposes of conducting a genetic study (i.e., data were not always collected at diagnosis stage). Third, since this sample was ascertained for a genetic study, there may be variation in severity of ASD symptoms compared to samples ascertained for other reasons. Fourth, this study was limited by the number and type of subgroups that could be evaluated to validate the phenotype. For example, it was not possible to examine the stability of the phenotype across IQ levels because of the different IQ measures used across the AGP data collection sites. Fifth, six (out of the 33) ADI-R algorithm items not relevant to non-verbal children were excluded from these analyses (e.g. reciprocal conversation and social chat) and so

the 6-factor solution might not entirely capture the symptom phenotype of verbal individuals. Another limitation is that splitting the sample by median age means there is the potential for measurement equivalence to differ by other categories of age, masked by the blunt division at the median in our sample. Finally, because the data on symptoms included are based on DSM-IV-TR it is possible that we inadvertently excluded high functioning girls (see Dworzynski, Ronald, Bolton & Happé, 2012). Therefore, if only more severely affected girls were included (in this sample) then group differences in factor means would not represent true sex differences.

Conclusions

Findings from this study suggest that individual items of the ADI-R algorithm, measure the same six constructs in children and youth with ASD in this study, regardless of verbal ability, sex, or age. Therefore, comparisons between pairs of subgroups (i.e., sub-groups associated with differences in the number of boys versus girls, verbal versus nonverbal status, and younger versus older) can be made on both latent construct and subgroup means. Furthermore, we can conclude that any potential differences on the latent factors of the ASD symptom phenotype would in fact represent true differences between these pairs of subgroups. This may not be true when using the two DSM 5 factors that while providing a good fit to the sample as a whole, are more unstable when comparing sub-groups.

It is important to note that even though the constructs measured by the items of the ADI-R have the same "meaning" in all three pairs of subgroups there are differences in "levels" of the latent factors of the autism phenotype between the verbal ability, sex and age subgroups in the AGP sample. Thus, the use of the ADI-R phenotypic structure in research should take into account the documented differences in symptom levels between these pairs of subgroups, such as age, sex, and verbal status in for example, genetic linkage studies (see Liu et al, 2011).

References

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders, fourth edition, text revision*. Washington DC: American Psychiatric Association.

American Psychiatric Association. (2010). Report of the DSM 5 Neurodevelopmental Disorders Work Group. http://www.dsm5.org. Accessed December 10, 2011.

Baumgartner, H., & Steenkamp, J-B. E. M. (1998). Multi-group latent variable models for varying numbers of items and factors with cross-national and longitudinal applications. *Marketing Letters*, *9*, 21–35.

Beckstead, J. W., Yang, C. Y., & Lengacher, C. A. (2008). Assessing cross-cultural validity of scales: A methodological review and illustrative example. *International Journal of Nursing Studies*, *45*(1), 110-119.

Bölte, S., & Poustka, F. (2001): [Factor structure of the Autism Diagnostic Interview-Revised (ADI-R): a study of dimensional versus categorical classification of autistic disorders]. *Z Kinder Jugendpsychiatr Psychother 2*9:221–9.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425– 440.

Byrne, B. M., Shavelson, R. J., & Muthén B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456-466. .

Byrne, B. M., & van de Vijver, F. J. R. (2010). Testing for Measurement and Structural Equivalence in Large-Scale Cross-Cultural Studies: Addressing the Issue of Nonequivalence. *International Journal of Testing*, 10(2), 107-132.

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

Dworzynski, K., Ronald, A., Bolton, P., & Happé, F. (2012). How different are girls and boys above and below the diagnostic threshold for autism spectrum disorders? *Journal of the American Academy of Child and Adolescent Psychiatry*, *51*(8), 788-797.

Frazier, T. W., Youngstrom, E. A., Kubu, C. S., Sinclair, L., & Rezai, A. (2008). Exploratory and confirmatory factor analysis of the autism diagnostic interview-revised. *Journal of Autism and Developmental Disorders*, *38*, 474–480.

Georgiades, S., Szatmari, P., Zwaigenbaum, L., Duku, E., Bryson, S., Roberts, W., Goldberg, J., & Mahoney, W. (2007). Structure of the autism symptom phenotype: A proposed multidimensional model. *Journal of the American Academy of Child and Adolescent Psychiatry*, *46*(2), 188–196.

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care. 44*, S78–S94.

Hoffman, E. J. (2009). Clinical features and diagnosis of autism and other pervasive developmental disorders. *Primary Psychiatry*, *16*(1), 36-44.

Hu, L., & Bentler, P. M. (1999). Cut-off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6* (1), 1–55.

Hus, V., Pickles, A., Cook, E. H., Jr, Risi, S., & Lord, C. (2007). Using the Autism Diagnostic Interview-Revised to increase phenotypic homogeneity in genetic studies of autism. *Biological Psychiatry*, *61*(4), 438–448.

Joreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*(4), 409-426.

Kamp-Becker, I., Ghahreman, M., Smidt, J., & Remschmidt, H. (2009). Dimensional structure of the autism phenotype: Relations between early development and current presentation. *Journal of Autism and Developmental Disorders*, *39*(4), 557–571.

Kenner, L. (1943). Autistic disturbances of affective contact. *Nervous Child*, *2*, 217-250.

Knight, G. P., & Zerr, A. A. (2010). Informed theory and measurement equivalence in child development research. *Child Development Perspectives*, *4*(1), 25-30.

Lecavalier, L., Aman, M. G., Scahill, L., McDougle, C. J., McCracken, J. T., Vitiello, B., Tierney, E., McCracken, J. T., Arnold, L. E., Ghuman, J. K., Posey, & D. J., Koenig, K. (2006). Validity of the Autism Diagnostic Interview–Revised. *American Journal on Mental Retardation*, *111*(3), 199–215.

Liu, X-Q., Georgiades, S., Duku, E., Thompson, A., Devlin, B., Cook, E.H., Wijsman, E.M., Paterson, A.D., Szatmari, P. (2011). Identification of genetic Loci underlying the phenotypic constructs of autism spectrum disorders. *Journal of the American Academy of Child and Adolescent Psychiatry*. *50*(7):687-696.e13

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

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B., DiLavore, P. C., Pickles, A., & Rutter, M. (2000). The Autism Diagnostic Observation Schedule-Generic: A standard

measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders, 30*, 205-233.

Lubke, G. H., & Muthén, B. (2004). Applying multigroup confirmatory factor models for continuous outcomes to likert scale data complicates meaningful group comparisons. *Structural Equation Modeling*, *11*(4), 514-534.

Millsap, R. E. & Kwok, O. M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations**.** *Psychological Methods* 4, *9*(1), 93-115.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*(3), 479–515.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115-132.

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*(4), 557-585.

Muthén, B., & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. *Mplus Web Notes: No. 4*. Retrieved 22/04/2011, from *www.statmodel.com*.

Muthén B., du Toit, S.H.C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes.* Unpublished manuscript.

Muthén, L. K. & Muthén, B. (2008). *Mplus 5.1 for Windows*. Los Angeles, CA: Author.

Prior, M., Eisenmajer, R., Leekam, S., Wing, L., Gould, J., Ong, B., & Dowe, D. (1998). Are there subgroups within the autistic spectrum? A cluster analysis of a group of children

with autistic spectrum disorders. *Journal of Child Psychology and Psychiatry*, *39*(6), 893-902.

Reise, S. P., Widaman, K. F., & Pugh, R. E. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*, 552-566.

Rutter, M., Le Couteur, A., Lord, C. (2003). *Autism Diagnostic Interview-Revised*. Los Angeles, USA: Western Psychological Services.

Risi, S., Lord, C., Gotham, K., Corsello, C., Chrysler, C., Szatmari, P., Cook, E. H. Jr, Leventhal, B. L., & Pickles, A. (2006). Combining information from multiple sources in the diagnosis of autism spectrum disorders. *Journal of the American Academy of Child and Adolescent Psychiatry, 45*(9), 1094-1103.

Snow, A. V., Lecavalier, L. A., & Houts, C. (2009). The structure of the autism diagnostic interview-revised: Diagnostic and phenotypic implications. *Journal of Child Psychology and Psychiatry*, *50*(6), 734–742.

Szatmari, P., Merette, C., Bryson, S., Thivierge, J., Roy, M.-A., Cayer, M., Maziade, M. (2002). Quantifying dimensions in autism: A factor analytic study. *Journal of the American Academy of Child and Adolescent Psychiatry*, *41*(4), 467–474.

Szatmari, P., Liu, X. Q., Goldberg, J., Zwaigenbaum, L., Paterson, A. D., Woodbury-Smith, M., Georgiades, S., Duku. E., & Thompson, A. (2012). Sex differences in repetitive stereotyped behaviors in autism: Implications for genetic liability. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *159B*(1), 5–12.

Tadevosyan-Leyfer, O., Dowd, M., Mankoski, R., Winkliskt, B., Putnam, S., McGrath, L., Tager-Flusberg, H., & Folstein, S. E. (2003). A principal components analysis of the

autism diagnostic interview-revised. *Journal of the American Academy of Child and Adolescent Psychiatry*, *42*(7), 864–872.

Vandenberg, R. J., Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4-69.

Van Lang, N., Boomsma, A., Sytema, S., Bildt, A., Kraijer, D., Ketelaars, C., & Minderaa, R. B.. (2006). Structural equation analysis of a hypothesised symptom model in the autism spectrum. *Journal of Child Psychology and Psychiatry*, *47*, 37–44.

Zwaigenbaum, L. (2010). Advances in the early detection of autism. *Current Opinion in Neurology*, *23*(2), 97-102.

Table 1. Results of categorical confirmatory factor analyses of competing models of the autism phenotype using items from the ADI-R

Model	χ^2(d.f.) = statistic	CFI	TLI	RMSEA
One-factor	$\chi^2(223)=8598.977$	0.745	0.907	0.102
Two factor	$\chi^2(216)=7174.292$	0.788	0.920	0.094
Three-factor DSM-IV	$\chi^2(218)=6407.919$	0.811	0.930	0.088
Six-factor (Liu et al., 2011)	**$\chi^2(220)=2286.084$**	**0.937**	**0.977**	**0.051**
2nd-order DSM-IV	$\chi^2(224)=2463.441$	0.932	0.975	0.052

Bolded is the final selected model

CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean square error of approximation

Table 2. Results of categorical confirmatory factor analyses of the selected measurement models of the autism phenotype within subgroups children with ASD

(Entries are χ^2(d.f.) = statistic, CFI, TLI, RMSEA)

	Selected model	2nd-order DSM-5 model
Verbal ability		
Verbal	χ^2(212)=1452.165 0.950, 0.979, 0.046	χ^2(211)=1464.710 0.949, 0.979, 0.047
Non-verbal	χ^2(187)=531.535 0.945, 0.973, 0.045	χ^2(189)=519.202 0.947, 0.974, 0.044
Sex		
Boys	χ^2(219)=2008.154 0.934, 0.976, 0.052	χ^2(212)=2536.408 0.963, 0.958, 0.046
Girls	χ^2(125)=301.642 0.963, 0.980, 0.050	n/a
Age group		
<88 months	χ^2(191)=1181.672 0.936, 0.972, 0.053	χ^2(194)=1180.740 0.936, 0.972, 0.053
>= 88 months	χ^2(208)=1119.343 0.945, 0.979, 0.049	n/a

CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean square error of approximation; n/a = no convergence

Table 3. Measurement equivalence across subgroups based on final selected model

Model	χ^2(d.f.) = statistic	DIFFTEST χ^2(d.f.) = statistic	CFI	TLI	RMSEA
Verbal ability (verbal vs. nonverbal)					
Unconstrained	χ^2(396)=1906.211		0.952	0.978	0.046
Constrained 1	χ^2(373)=2281.555	χ^2(43)=529.584	0.939	0.971	0.053
Constrained 2	χ^2(376)=2104.334	χ^2(40)=390.452	0.945	0.974	0.050
Sex (boys vs. girls)					
Unconstrained	χ^2(319)=1845.801		0.949	0.977	0.051
Constrained	χ^2(321)=1734.270	χ^2(47)=156.281	0.953	0.978	0.049
Age group (<88m vs. >= 88m)					
Unconstrained	χ^2(398)=2304.195		0.940	0.976	0.051
Constrained	χ^2(387)=2282.742	χ^2(46)=296.559	0.940	0.975	0.052

CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean square error of approximation

Table 4. Descriptive statistics of AGP sample based on factors of final selected measurement model for the autism phenotype

(Entries are mean, SD)

	Factors of the autism phenotype					
	JTATT	SOCINT	NVCOMM	PEERINT	RSMB	CRINT
Verbal ability						
Nonverbal	1.89, 0.49	2.31,0.49	1.80, 0.53	2.11, 0.61	1.65, 0.55	0.80, 0.65
Verbal	1.56, 0.58	1.87, 0.60	1.27, 0.76	1.78, 0.71	1.13, 0.59	1.05, 0.68
Effect Size	**0.57**	**0.73**	**0.70**	**0.47**	**0.82**	**-0.37**
Sex						
Boys	1.64, 0.58	1.98, 0.60	1.69, 0.75	1.86, 0.70	1.27, 0.62	1.00, 0.68
Girls	1.67, 0.59	1.97, 0.64	1.49, 0.74	1.85, 0.72	1.25, 0.63	0.88, 0.65
Effect Size	-0.06	0.02	**0.13**	0.03	0.04	**0.18**
Age group						
<=88 months	1.51, 0.57	0.87, 0.60	1.37, 0.73	1.77, 0.69	1.33, 0.69	0.88, 0.65
> 88months	1.78, 0.56	2.09, 0.59	1.44, 0.76	1.95, 0.70	1.21, 0.64	1.08, 0.69
Effect Size	**0.47**	**0.36**	**0.10**	**0.26**	**-0.19**	**0.30**

Bolded effect sizes are statistically significant at p<0.05 using ANOVA

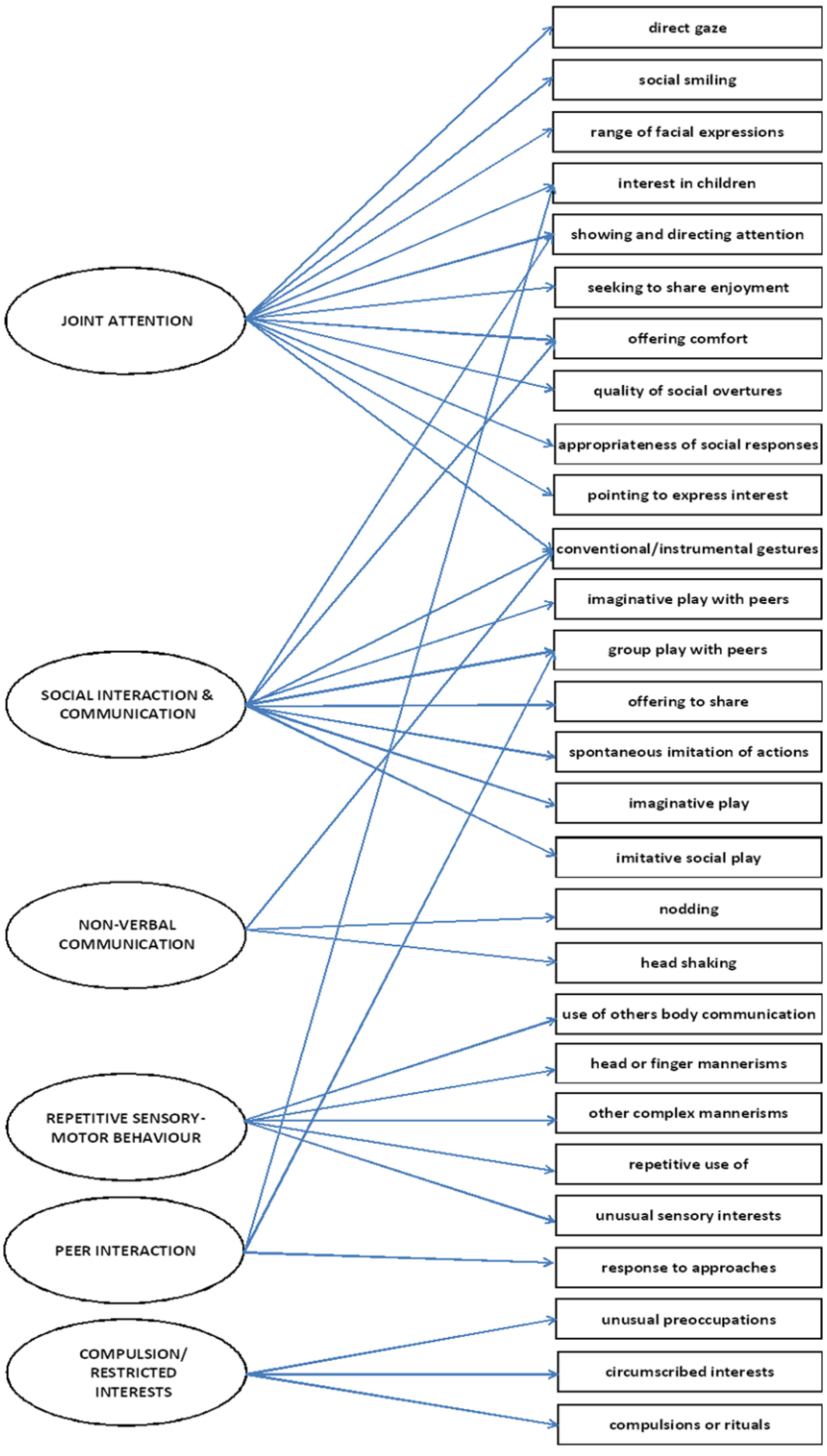
Legend:

JTATT = joint attention; SOCINT = social interaction & communication;

NVCOMM = non-verbal communication; PEERINT = peer interaction;

RSMB = repetitive sensory-motor behaviour; CRINT = compulsion/ restricted interests.

Figure 1. Selected first-order measurement model of the Autism phenotype



Latent Factors ADI-R algorithm items

Appendix 1. Second-order measurement model based on proposed DSM 5 conceptualization

2nd order latent factors 1st order latent factors ADI-R algorithm items

Appendix 2. Descriptive statistics overall and internal consistencies of 6-factor model

	Mean	Standard Deviation	Cronbach's α
JTATT	1.64	0.58	0.86
SOCINT	1.98	0.61	0.85
NVCOMM	1.41	0.75	0.77
PEERINT	1.86	0.71	0.75
RSMB	1.28	0.64	0.56
CRINT	0.98	0.68	0.30[2]

Legend:

JTATT = joint attention; SOCINT = social interaction & communication;

NVCOMM = non-verbal communication; PEERINT = peer interaction;

RSMB = repetitive sensory-motor behaviour; CRINT = compulsion/ restricted interests.

[2] The low Cronbach's alpha is an indication of the lack of variability or interconnectedness of the items in the compulsion restricted interests construct.

Chapter 4. Validation of the BRIEF-P in a sample of Canadian preschool children

Validation of the BRIEF-P in a sample of Canadian preschool children

Eric Duku[1,2] and Tracy Vaillancourt[1,2]

[1]University of Ottawa, Ottawa, ON, Canada; [2]Offord Centre for Child Studies, McMaster University, Hamilton, ON, Canada.

Abstract

The Behavior Rating Inventory for Executive Function-Preschool (BRIEF-P) is an instrument designed to assess preschoolers' executive function (EF) in the context of where the behaviour occurs. The present study examined the psychometric properties and measurement structure of the BRIEF-P using parents' and teachers' reports on 625 children aged between 25 and 74 months. Results indicated that the BRIEF-P scales had good internal consistency and convergent validity in this sample of children. However, the measurement models examined exhibited poor fit statistics and showed that the EF construct was not unidimensional but rather multidimensional with interrelated sub-constructs. Further analyses showed that three of the clinical scales (Emotional Control, Plan/Organize and Working Memory) were unidimensional and invariant across informant. The other two clinical scales (Inhibit and Shift) were multidimensional and differed by informant. Results support a multidimensional construct of EF and accordingly, different measurement models are proposed by informant.

Keywords: executive function; validation; preschool, confirmatory factor analysis; invariance; informant

Validation of the BRIEF-P in a sample of Canadian preschool children

Executive functioning (EF) describes a set of cognitive abilities that control and regulate other abilities and behaviour (Gioia, Isquith, Kenworthy & Barton, 2002). EF is necessary for goal-directed behaviour and includes the ability to start and stop actions, to monitor and change behaviour as needed, and to plan future behaviour when faced with novel tasks and situations. EF permits the anticipation of outcomes and the adaption to changing situations. As well, the ability to form concepts, and the ability to think abstractly, are often considered components of EF (Gioia et al 2002; Isquith, Crawford, Espy & Gioia, 2005; Mahone et al, 2002).

There has been some debate in the research literature concerning whether EF is a unitary construct or a multidimensional construct with interrelated yet distinct processes (Jurado & Rosselli, 2007; Senn, Espy, & Kaufmann, 2004). For example, Duncan et al. (1996) have argued that there is a common mechanism and a central factor underlying EF. Conversely, Miyake et al. (2000) have argued there are three components to EF – Shifting (ability to switch between different tasks), Updating (continuous monitoring and quick addition or deletion of contents within an individual's working memory), and Inhibition (capacity to override responses that are influential in a given situation) that are distinguishable with some underlying commonality.

The Behavior Rating Inventory of Executive Function – Preschool (BRIEF-P; Gioia, Espy & Isquith, 2003) is an instrument that is purported to measure EF in children aged two years to five years and eleven months. The BRIEF-P is completed by parents, teachers and or caregivers in the context of where the child's behaviour (that is being assessed) occurs (e.g., home, daycare). The instrument was developed conceptually so that the items in each scale

reflected the intended domain of EF. The domains of EF were defined using theory, clinical practice, and research literature (Gioia et al., 2003).

The BRIEF-P has sixty-three items and the following five non-overlapping clinical scales: (1) Inhibit (16 item), (2) Shift (10 items), (3) Emotional Control (10 items), (4) Working Memory (17 items), and (5) Plan/Organize (10 items). The *Inhibit* scale measures the child's ability to inhibit, resist or not act on an impulse and the ability to stop his or her own behaviour at the appropriate time. The *Shift* scale measures the child's ability to move freely from one situation, activity, or aspect of a problem to another, as the circumstances demand. The *Emotional Control* scale addresses the manifestation of EF within the emotional realm and measures a child's ability to modulate emotional responses. The *Working Memory* scale measures the child's capacity to hold information in mind for the purpose of completing a task or making a response and is necessary for carrying out multi-step activities, activities, implementing a sequence of actions or following complex instructions. The *Plan/Organize* scale measures the ability of the child to manage current and future-oriented task demands within the situational context. The *Plan* component relates to the ability to anticipate future events, implement instructions or goals and develop appropriate steps ahead of time in order to carry out a task or activity. The *Organize* component of this scale relates to the ability to bring order to information, actions or materials to achieve and objective.

The five clinical scales can be summarized into three overlapping broader indices: (1) Inhibitory Self-control, (2) Flexibility and (3) Emergent Meta-Cognition. An overall composite score termed the Global Executive score can also be computed (see Figure 1 for schematic of BRIEF-P). Higher scores on any of the clinical scales are indications of poorer EF.

Although the BRIEF-P is used extensively, particularly in clinical settings, to assess differences in subgroups of children, its measurement properties have only been examined in two studies (Gioia et al., 2003; Bonillo et al., 2011). The limited evaluation of the measurement properties of the BRIEF-P is problematic in that it is unclear if the measure functions the same way across different informants (i.e., measurement equivalence). Since the BRIEF-P was intended to be completed by both parents and teachers/childcare workers, establishing measurement invariance across informants is necessary for this tool (Gioia et al., 2003).

Establishment of measurement equivalence is a fundamental component of research. A central principle of measurement equivalence purports that the scores or ratings across different groups are on the same scale and that the empirical relationships between test items and traits of interest remain stable across groups (Reise, Widaman, & Pugh, 1993). Establishing the measurement equivalence of an instrument across groups and subgroups within a population helps support the degree of generalizability of the constructs under study (Prince, 2008). Indeed, constructs need to be replicable and generalizable beyond the group (e.g., country) for which they were developed. In establishing measurement equivalence, a detailed study of the psychometric properties of the construct, including examination of convergent validity, is needed. More importantly, the impact of characteristics such as type of informant on measurement invariance must be assessed.

The two published studies examining the measurement properties of the BRIEF-P have reported some similar *and* discrepant findings. In the first study, Gioia et al. (2003) found that a 3-factor solution best fit the data obtained from both teacher (n=302) and parent informants (n=460), using data for normative samples of preschoolers aged two to five years. The analytic strategy used by Gioia et al. included conducting a series of exploratory factor analyses (EFA)

with promax rotation using the five clinical scale scores instead of the individual items. Using scale scores and not items in an EFA is called *item parcelling*. Item parcelling is a common practice among researchers, however is not a recommended unless certain important conditions have been met, such as first establishing the unidimensionality of the scale scores (Bandalos, 2002; Little et al., 2002). In fact, researchers are discouraged from using item parcels as indicators in tests of measurement equivalence because this process has been shown to affect tests of equal factor loadings across groups (Meade & Kroustalis, 2006). Specifically, differences in scores between groups are masked if problems with items are not seen, thus leading researchers to accept the hypothesis of equivalence of models between groups when in fact equivalence is not present. Results from Gioia et al.'s EFA using item parcelling indicated three factors: (1) Emergent Meta-Cognition, which included the scales Working Memory, Plan/Organize, and Inhibit, (2) Flexibility, which included the scales Shift and Emotional Control, and (3) Inhibitory Self-Control, which included the scales Emotional Control and Inhibit.

In the second study, Bonillo et al. (2011) validated the Catalan version of the BRIEF-P using a cluster random sample of children aged 3 to 6 years recruiting 417 teachers and 408 parents as informants. Using an EFA with Promax rotation Bonillo et al. showed that the Spanish version had excellent reliability and demonstrated that the three broader indices of Inhibitory Self-control, Flexibility and Emergent Meta-cognition fit the data as depicted by Gioia et al. (2003; see also Isquith et al., 2004). In this case, item parcelling was also used. However the authors also attempted to conduct a confirmatory factor analysis (CFA) using the 63 items. CFA is a statistical technique used to verify the existence of a measurement model for a set of observed variables. Specifically, the CFA approach evaluates the hypothesis that the relationship

between observed variables and their underlying latent construct(s) exists as stated in the measurement model. In this case, the results of the CFA indicated that correlations between the latent constructs were greater than one which is evidence of a misspecified model since it is impossible for the correlation to be greater than one (Kline, 2011) and hence the test of fit of the hypothesized model suggested by the developers could not be undertaken.

Present Study

As mentioned, the BRIEF-P is widely used in research and in clinical settings. However, very few studies have been conducted on its measurement properties, and the two studies that have, examined the fit of the measurement model without item parcelling and formally testing the equivalence of the measurement across informants (e.g., Bonillo et al., 2011; Gioia et al, 2003). Given our limited understanding of the BRIEF-P's measurement properties, the purpose of this study was to comprehensively examine its psychometric properties, measurement model, and the invariance of the model across informants. Toward this aim, data from Canadian preschool children were used along with a concurrent child outcome measure, the Child Behavior Checklist 1.5-5 (Achenbach & Rescorla, 2000; CBCL henceforth).

To replicate the earlier work by Gioia et al. (2003) and Bonillo et al. (2011), CBCL scales were also used to establish convergent validity. Since EF affects behaviour, it can be assumed that some of the subscales within the CBCL should be related or associated with the clinical scales from the BRIEF-P. For example, Working Memory is thought to contribute to attention disorders and should be associated with the Attention Problems subscales of the CBCL (Gioia et al., 2003). Similarly, the Inhibit scale should be associated with the Attention Problems subscale of the CBCL and correlated less with the Anxiety and Somatic Complaints subscales of the CBCL.

Method

Participants

Data on 625 children aged from 25 months to 74 months were used in this study. The criteria for inclusion in the study were that the child being reported on had to: (1) be registered in a licensed childcare facility in the community, and (2) not have a known developmental delay. Data for the BRIEF-P (Gioia et al., 2003) were obtained from 60 licensed childcare facilities from both parents (n=479) and early childhood educators (teachers; n=606). Data on known correlates of executive functioning were collected using the CBCL. As well, socio-demographic data such as sex of the child and age in months of the child were also collected.

Measures

Behaviour Rating Inventory Executive Functioning – Preschool Version (BRIEF-P). The BRIEF-P is a questionnaire for parents and caregiver/teachers of children who are of pre-school age. It was designed to assess EF behaviour and is intended for a broad age range of children, 2 years through 5 years 11 months. The goal of the developers was to create an instrument that has good internal consistency and validity appropriate to a behavior rating scale (Gioia et al, 2003). All items are scored on a scale of 1 (never), 2 (sometimes), and 3 (always). The developers report good internal consistency (0.80 to 0.97) and test-retest reliability (0.65 to 0.94). Inter-rater reliability was reported to be between a low of 0.06 for the Plan/Organize scale to 0.28 for the Shift scale.

Child Behavior Checklist 1.5-5 (Achenbach & Rescorla, 2000). The 99-item CBCL is a widely-used norm-referenced instrument that can evaluate a wide range of internalizing and externalizing disorders, based on six subscales (Emotionally Reactive, Anxious/Depressed, Somatic Complaints, Withdrawn, Attention Problems, Aggressive Behaviour; Achenbach &

Rescorla, 2000). The CBCL is completed by primary caregivers such as parents or teachers based on observation of the child's behaviour in the previous two months. Scale and subscale scores are summed and converted to T-scores, with scores greater than 70 considered to be within the "clinically significant" range. The CBCL has good reported test-retest reliability over a period of a week (0.68-0.92) and inter-rater reliability (0.48-0.67) for all scales and subscales (Achenbach & Rescorla, 2000). The developers report evidence of discriminative, convergent, and predictive validities.

Analytic Plan

Data were first examined to determine if there were any differences by age, sex of children, and missing data patterns by number of informants (one versus two) to ensure there was no bias in scores by number of informants who rated a child. Next, the psychometric properties of the BRIEF-P were examined by replicating the earlier analyses by the developers. This comprised of examining the (a) internal consistency of the clinical scales; (b) convergent validity of the clinical scales with the CBCL subscales by examining the correlations between the clinical scales of the BRIEF-P and the subscales of the CBCL; (c) inter-informant reliability, and (d) goodness of fit of the measurement model suggested by the developers of the BRIEF-P. Assessment of the second-order measurement model (using scales scores) suggested by the developers of the BRIEF-P (see Figure 1) was not possible because the fit could not be evaluated. Specifically, the models were "not identified" because each of the factors in the measurement model had only two indicators since one of the conditions for model identification is that latent factors have a minimum of three indicators (Kline, 2011). Finally, the goodness of fit of competing models for the EF construct and the unidimensionality of the measurement

models for each of the clinical scales were tested using Categorical Confirmatory Factor Analysis (CCFA).

If the measurement model of any of the clinical scales did not meet the criteria for unidimensionality, two subsamples of equal numbers of randomly selected children were drawn to derive new measurement models for the non-unidimensional scales. In this case, one subsample was assigned to the Categorical Exploratory Factor Analysis (CEFA) sample and the other to the Categorical Confirmatory Factor Analysis (CCFA) sample which functioned as the validation subsample. This analytic approach allowed us to validate the model structure of the CEFA with CCFA and it allowed us to compare the measurement properties by informant (parents and teachers). Model selection for the CEFA was based on parsimony, examination of the scree plot, the value of the Root Mean Square Error of Approximation (RMSEA), and the criterion that at least three of the loadings for a factor be 0.3 or higher and allowing items to cross-load. Hu and Bentler (1999) suggest using evidence of model goodness of fit based on findings from multiple indices, and for this study, we used the Tucker Lewis Index (TLI), Comparative Fit Index (CFI) and the RMSEA to evaluate model fit. For the CCFA (validation subsample) criteria for acceptance of fit of the measurement model was based on the same three goodness of fit criteria: the TLI, CFI, and the RMSEA.

The RMSEA is an estimate of how well the fitted model approximates the population covariance matrix per degrees of freedom. Hu and Bentler (1999) recommend a cut-off value of .06 (although 0.08 is considered to be adequate and the upper limit for the RMSEA) and using a p-value for testing the hypothesis that the discrepancy is smaller than .05. The TLI and CFI measure the proportionate improvement in fit between the baseline model and the target model, per degrees of freedom and Hu and Bentler recommend using a cut-off value of .95. However, a

cutoff of 0.92 has been shown to be acceptable (Byrne & van de Vijver, 2010). All three indices are not as affected by sample size as are other measures of model fit (Hu & Bentler, 1999).

If any of the clinical scales met the criteria for unidimensionality, we examined the invariance of their measurement models across informant within a structural equation modeling framework using multiple group categorical confirmatory factor analyses (MGCCFA) in Mplus (Muthén, 1984; Muthén, du Toit, & Spisic, 1997). Following recommendations by Muthén and Asparhouhov (2002), the MGCCFA analysis was conducted as follows. First, thresholds and factor loadings were unconstrained (or freed) across both informants, while the scale factors were constrained (fixed) to one and factor means were constrained to zero across both informants. Second, thresholds and factor loadings were constrained to be equal across the two informants with the scale factors fixed at one for one informant and freed in the other, and factor means fixed at zero in one group of informants and freed in the other group. A difference of chi-square test was computed where a statistically significant chi-square meant that the hypothesis that the constrained parameters were invariant across groups (i.e., the equality constraints are violated) was rejected. Here, the criterion that changes in CFI and TLI was less than 0.01 when comparing the baseline model with the models where parameters were not constrained was used as evidence of invariance (Byrne & van de Vijver, 2010). If measurement invariance was not found, partial measurement invariance was evaluated by relaxing some of the constraints with the restricted model (Muthén & Muthén, 2008).

Results

The BRIEF-P endorsement rates of items by informants were over 5%, thus ensuring that all items could be used in examination of the measurement models of the EF construct. The number of informants rating a child (one versus two informants) did not differ by the child's sex

(44% girls versus 49% girls; Fisher exact test: $p = 0.28$) nor by the child's age (mean age: 39.0 months versus 39.1 months; F (1,621) = 0.58, MSE = 62.38; $p = 0.45$).

Internal consistency and convergent validity of the clinical scales

Table 1 presents the internal consistencies of the clinical scales by informant. All of the clinical scales had internal consistencies over 0.8 indicating that the items within each scale measure similar constructs. The internal consistencies for the clinical scales based on teacher report were always higher than the internal consistencies based on parent report. The inter-informant (parent-teacher) reliabilities for the sample are also presented in Table 1. Although the reliability estimates for the clinical scales are low, the correlations between the scales (r=0.16 to 0.38) were similar in magnitude to those reported by the developers of the BRIEF-P (Gioia et al., 2003).

Convergent validity was examined by comparing the correlations between the clinical scales of the BRIEF-P and the subscales of the CBCL. The correlations between the BRIEF-P clinical scales and the CBCL subscales by informant are presented in Table 2. The correlations were low to moderate, varying from r=0.15 (Plan/Organize and Somatic Complaints – teacher ratings) to r=0.81 (Inhibit and Aggressive behaviour – teacher ratings). The correlations between the BRIEF-P scales and the CBCL subscales were of similar magnitude and direction to those reported by the developers of the BRIEF-P (Gioia et al., 2003). As expected, the Working Memory was associated with the Attention Problems subscale of the CBCL (parent r=0.63, teacher r=0.82). Similarly the Inhibit scale was more strongly associated with the Attention Problems subscale of the CBCL (parent r=0.68, teacher r=0.81) than with the Somatic Complaints subscale (parent r=0.25, teacher r=0.15). Finally, the Emotional Control clinical

scale was correlated with the Emotional Reactivity CBCL subscale (parent r=0.61, teacher r=0.76)

Measurement models for the BRIEF-P overall

As mentioned earlier, the second-order measurement model suggested by the BRIEF-P developers (see Figure 1), could not be examined because the models were "not identified" given that each of the factors in the measurement model had only two indicators. Two competing measurement models: (1) unidimensional (or unitary) 63-item model and (2) first-order five factor 63-item model (non-unitary) were examined by informant. The results for the competing measurement models are presented in Table 3. Results indicated that the fit statistics for both the unidimensional model and the five-factor model did not meet conventional fit statistics criteria.

Measurement models for the clinical scales

Since the BRIEF-P was developed conceptually scale by scale, the unidimensionality of the measurement models for each of the individual clinical scales was evaluated using CCFA. The results of the CCFA analyses are presented in Table 4. Results indicated that the measurement models for the Emotional Control, Plan/Organize and the Working Memory clinical scales met the criteria for unidimensionality for parent and teacher reports. However, the two clinical scales, Inhibit and Shift, did not meet the criteria for unidimensionality based on both parent and teacher reports.

Given the poor fit of the Inhibit and Shift clinical scales, we re-examined their measurement properties using Categorical EFA by informant using the EFA subsample. The factor loadings of the final selected measurement models for the Inhibit and Shift clinical scales by informant are shown in Figure 2. There were differences in the final derived/proposed measurement models for inhibit and shift scales by informant using this statistical approach.

From Figure 2, we can see the derived Inhibit scales for parent reports have items that are subsets of the items in the derived Inhibit scales for teacher reports. For example, the Inhibit scale I for teacher reports has four items (BRIEF 13, 23, 28, and 58). The Inhibit scale II for teacher reports has two items (BRIEF 43 and 46) more than the Inhibit scale II for parent reports. Similar differences are seen in the Shift scales I, II and III by informant.

The CCFAs for the final selected measurement models by report were acceptable. The results were as follows: (1) the inhibit scale (Inhibit I and Inhibit II scales combined) - $\chi^2(47) = 79.196$; CFI = 0.977, TLI = 0.990, RMSEA=0.053 for parent reports and $\chi^2(42) = 112.660$; CFI = 0.984, TLI = 0.995, RMSEA = 0.074 for teacher reports; and (2) the Shift scale (Shift I, Shift II and Shift III scales combined) - $\chi^2(20) = 47.707$; CFI = 0.968, TLI = 0.981, RMSEA = 0.075 for parent reports, and $\chi^2(15) = 35.836$; CFI = 0.988, TLI = 0.993, RMSEA = 0.068 for teacher reports. The internal consistencies for the derived scales were all above 0.6 (see Table 5).

Measurement invariance of the clinical scales by informant

Based on our criteria for assessing measurement invariance across informants, two of the three unidimensional clinical scales, Plan/Organize (DIFFTEST: $\chi^2(14) = 22.768$; ΔCFI = 0.002, ΔTLI = 0.003, ΔRMSEA = -0.008) and Working Memory (DIFFTEST: $\chi^2(22) = 79.193$; ΔCFI = -0.001, ΔTLI = 0.000, ΔRMSEA = 0.000), were invariant across informants (see Table 6). The Emotional Control clinical scale was partially invariant (DIFFTEST: $\chi^2(11) = 64.351$; ΔCFI = 0.007, ΔTLI = -0.002, ΔRMSEA = 0.011) and this was achieved by allowing the item "*reacts more strongly to situations than other children*" to be unconstrained across informants.

Proposed Executive Function Measurement Model based on the BRIEF-P

Finally, the measurement models derived for the Shift and Inhibit clinical scales were then integrated with the measurement models for the other three clinical scales into overarching

final and proposed measurement models by informant for the EF construct (see Figures 3 and 4). The proposed measurement model for the EF construct differed from the original proposed by the developers of the BRIEF. For example, instead of a single construct for the Inhibit clinical scale, the proposed measurement model comprises two latent constructs measuring the Inhibit scale with some cross-loading items. For the Shift clinical scale, the proposed measurement model consists of three latent constructs with cross-loading items instead of a single latent construct.

Discussion

This is the first study to examine the invariance of the measurement model of EF using the BRIEF-P across informants with data collected on a community sample of preschool children. The psychometric properties and validity of the BRIEF-P in a community sample of preschool children were also examined. The present findings showed that the individual clinical scales had good internal consistency and convergent validity with a child measure of psychopathology (CBCL) consistent with the findings reported by the developers (Gioia et al., 2003). As evidence of convergent validity, the BRIEF-P was associated with measures of behaviour and attentional functioning measured by the CBCL, since items on the BRIEF-P measure EF through manifestations in behaviour (Isquith et al, 2005).

Investigation of the competing measurement models for EF allowed for generalization to the community or population at large. Our analyses of competing measurement models using the items of the BRIEF-P provided evidence that the EF construct was not unitary in nature but rather could be conceptualized as multidimensional with interconnected sub-dimensions.

Evaluation of the clinical scales showed that two of the clinical scales, Inhibit and Shift, were not unitary and that the Inhibit scale could be split into two scales while the Shift scale

could be split into three scales. All five scales had different factor structures by informant and had acceptable internal consistencies. Differences in measurement models for these scales by informant could be attributed to differences in context and behaviour observed by informants. There is also the added challenge of how to integrate data from different informants because of differences in context and low agreement in ratings between informants. It is important that we understand the reasons for differences in informant perception and differences in measurement models by informant since it is used for assessing executive function. These differences could be due to differential identification of children meeting criteria for disorder (De Los Reyes & Kazdin, 2005; Offord et al., 1996). That is, discrepancies among informants' rating could arise from the way informants weight differently the possible contextual causes and the possible causes of the observed behaviour exhibited by the child. It is therefore possible for the informant's rating to be influenced by contextual variables such as having an older sibling or to be based on cultural/social values. These differences require further investigation and replication in other samples. Dirks et al. (2012), suggest that investigators examine the differences in informant's ratings as meaningful clinical information and move toward approaches that embrace contextual variability in children's behaviour. The advantage of having multiple ratings per child is that a more complete picture of the child is available which can be used to revise the scales to take advantage of the context-dependent nature of the ratings.

Similarity of the measurement models (configural invariance) across informant was established for the Emotional Control, Plan/Organize and Working Memory clinical scales. Configural measurement invariance is the first step in the process of establishing measurement invariance (metric and scalar). A construct is assumed to be invariant when members of different populations who have the same characteristics and attributes receive the same observed

score (Teresi, 2006). Ensuring construct comparability between informants is important since lack of equivalence can jeopardize cross-informant comparisons. The MGCCFA approach used in this study can be extended to test equivalence of the measurement models of the clinical scales across subgroups of children. Multiple group comparisons can also help point out similarities and differences in the structure of the executive function construct across groups.

Using the eight scales based on the three original scales and the five integrated scales our findings suggest a non-unitary eight-dimensional measurement model of EF could be used, which is also different by informant (Figures 3 and 4). Our proposed measurement model is different from other analyses of the factor structure because our analyses were performed at the item level within the clinical scales. This is a better approach than using item parcels to examine the measurement model of the EF construct. The Inhibit scale had two subconstructs with some cross-loading items whereas the Shift scale comprised of three constructs with some cross-loading items. A larger independent sample with data from both informants is required to confirm the equivalence of the proposed measurement models across informant so as to not run into estimation problems.

Some limitations to this study need to be acknowledged. First is the challenge of assessing EF in preschool children based on observed performance by informant. Currently there are no guidelines that exist on what to do with discrepant information and how to interpret such discrepant information. Second is the limitation of using data from parents and teachers as informants and the possible influence of context bias on ratings. Third, all the children came from a childcare setting which means they had a shared experience that not all children enjoy and this may limit the generalizability of our findings to children in other care settings.

Conclusions

Based on the analyses of the BRIEF-P in this study we were able to demonstrate in a community sample of preschool age children that the BRIEF-P scales have good internal consistency and convergent validity. We also provided evidence that three of the clinical scales of Emotional Control, Plan/Organize and the Working Memory were invariant across informant. Finally, we proposed a multidimensional measurement model for EF that supports the theory that the EF construct is not unitary but, rather, is multidimensional with interrelated/interconnected dimensions. Further work is needed to replicate the results of this study in other studies with larger samples (in order to not run into estimation problems) and with samples in different settings (e.g., clinical). Further work is also needed on the clinical scales so that it makes use of the context-dependent nature of such ratings.

References

Achenbach, T. M., & Rescorla L. A. (2000). *Manual for the ASEBA preschool forms and profiles*. Burlington: University of Vermont.

Bandalos, D. (2002). The Effects of Item Parceling on Goodness-of-Fit and Parameter Estimate Bias in Structural Equation Modeling. *Structural Equation Modeling, 9*(1), 78–102.

Bonillo, A., Araujo Jiménez, E., A., Jané Ballabriga, M., C., Capdevila, C., & Riera, R. (2012). Brief Report: Validation of Catalan Version of BRIEF-P. *Child Neuropsychology, 18*(4), 347-55.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Byrne, B. M. & van de Vijver, F. J. R. (2010). Testing for Measurement and Structural Equivalence in Large-Scale Cross-Cultural Studies: Addressing the Issue of Nonequivalence. *International Journal of Testing*, *10*(2), 107-132.

De Los Reyes, A., & Kazdin, A.E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin, 131,* 483-509.

Dirks, M. A., De Los Reyes, A., Briggs-Gowan, M. Cella, D., & Wakschlag, L.S. (2012). Annual Research Review: Embracing not erasing contextual variability in children's behavior – theory and utility in the selection and use of methods and informants in developmental psychopathology. *Journal of Child Psychology and Psychiatry*, *53*(5): 558-574.

Donders, J. (2002). The behavior rating inventory of executive function. *Child Neuropsychology, 8*(4), 229-230.

Duncan, J., Emslie, H., Williams, P., Johnson, R., & Freer, C. (1996). Intelligence and the frontal lobe: The organization of goal-directed behavior. Cognitive Psychology, 30, 257-303.

Gioia, G. A., Espy, K. A., & Isquith, P. K. (2003). *Behavior Rating Inventory of Executive Function, Preschool Version (BRIEF-P)*. Odessa, FL: Psychological Assessment Resources.

Gioia, G. A., Isquith, P. K., Kenworthy, L., & Barton, R. M. (2002). Profiles of everyday executive function in acquired and developmental disorders. *Child Neuropsychology, 8*, 121–137.

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

Isquith, P, K, Gioia, G, & Espy, K, A. (2004). Executive function in preschool children: examination through everyday behavior. Developmental Neuropsychology, *26*, 403–422.

Isquith, P. K., Crawford, J. S., Espy, K.A., & Gioia, G.A. (2005). Assessment of executive function in pre-school aged children. *Mental Retardation and Developmental Disabilities Research Reviews, 11*, 209–215.

Jurado, M.B. & Rosselli, M (2007). The elusive nature of Executive functions: A review of our current understanding. *Neuropsychology Review*, 17(3), 213-233.

Kline, R.B. (2011). *Principles and Practice of Structural Equation Modeling, 3rd edition*. New York: Guilford Press.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling*, *9*, 151–173.

Mahone, E. M., Cirino, P. T., Cutting, L. E., Cerrone, P. M., Hagelthorn, K. M., et al. (2002). Validity of the behavior rating inventory of executive function in children with ADHD and/or Tourette syndrome. *Archives of Clinical Neuropsychology, 17*(7), 643–62.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41*, 49–100.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115-132.

Muthén, B., & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. *Mplus Web Notes: No. 4*. Retrieved 22/04/2011, from *www.statmodel.com*.

Muthén B., du Toit, S.H.C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes.* Unpublished manuscript.

Muthén, L. K. & Muthén, B. (2008). *Mplus 5.1 for Windows*. Los Angeles, CA: Author.

Offord, D. R., Boyle, M. H., Racine, Y., Szatmari, P., Fleming, J. E., Sanford, M., & Lipman, E. L. (1996). Integrating assessment data from multiple informants. *Journal of the American Academy of Child and Adolescent Psychiatry*, *35*, 1078-1085.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological Bulletin, 114*, 552-566.

Senn, T. E., Espy, K. A., & Kaufmann, P. M. (2004). Using path analysis to understand executive function organization in preschool children. *Developmental Neuropsychology, 26*(1), 445-464.

Sherman, E.M.S., & Brooks, B.L. (2010). Behavior Rating Inventory of Executive Function - Preschool Version (BRIEF-P): Test review and clinical guidelines for use. *Child Neuropsychology, 16*(5), 503-519.

Teresi, J. A. (2006). Overview of quantitative measurement methods equivalence, invariance, and differential item functioning in health applications. *Medical Care*, *44*(11), s39-s49.

Table 1. Internal consistencies of scales of the BRIEF-P by informant and inter-informant reliability

Scales	Internal consistency		Inter-informant reliability (n=458)
	Parent (n=479)	Teacher (n=606)	
Inhibit	0.889	0.949	0.385**
Shift	0.838	0.880	0.194**
Emotional control	0.857	0.922	0.164**
Working memory	0.903	0.945	0.312**
Plan/organize	0.799	0.898	0.232**

** $p<0.001$

Table 2. Convergent validity of the BRIEF-P scales using the CBCL 1.5-5 scales by informant

	CBCL 1.5-5 scales						
Parents (n=458)							
BRIEF scales	Emotional reactive	Anxious/ Depressed	Somatic Complaints	Withdrawn	Sleep problems	Attention problems	Aggressive behaviour
Inhibit	.44**	.36**	.25**	.44**	.35**	.68**	.70**
Shift	.59**	.52**	.41**	.45**	.23**	.32**	.42**
Emotional control	.61**	.52**	.35**	.47**	.28**	.40**	.66**
Working memory	.40**	.366**	.27**	.54**	.24**	.63**	.54**
Plan/organize	.37**	.36**	.22**	.50**	.32**	.59**	.58**
Teachers (n=579)							
BRIEF scales	Emotional reactive	Anxious/ Depressed	Somatic Complaints	Withdrawn		Attention problems	Aggressive behaviour
Inhibit	.43**	.20**	.15**	.45**		.81**	.82**
Shift	.67**	.63**	.32**	.53**		.30**	.34**
Emotional control	.76**	.57**	.29**	.48**		.52**	.74**
Working memory	.40**	.28**	.18**	.61**		.81**	.61**
Plan/organize	.36**	.25**	.15**	.59**		.77**	.58**

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

Table 3. Results of the CFA for the competing measurement models for Executive Function by informant

(a) Parent

	Unidimensional measurement model	1st-order 5 factor model
$\chi 2$(d.f.) = statistic	$\chi 2(186) = 1176.280$	$\chi 2(196) = 634.779$
CFI	0.715	0.847
TLI	0.902	0.959
RMSEA	0.105	0.068

(b) Teacher

	Unidimensional measurement model	1st-order 5 factor model
$\chi 2$(d.f.) = statistic	$\chi 2(125) = 307.3164$	$\chi 2(166) = 1352.866$
CFI	0.740	0.860
TLI	0.919	0.967
RMSEA	0.171	0.109

CFI comparative fit index; TLI = Tucker Lewis Index; RMSEA = Root Mean Square Error of Approximation.

Table 4. Results of test for unidimensionality of the clinical scales by informant

(a) Parent

	Inhibit	Shift	Emotional control	Working memory	Plan /organize
$\chi 2$(d.f.) = statistic	**$\chi 2$(53) = 379.775**	**$\chi 2$(23) = 235.828**	$\chi 2$(27) = 94.856	$\chi 2$(48) = 119.122	$\chi 2$(26) = 82.075
CFI	**0.879**	**0.866**	0.970	0.949	0.954
TLI	**0.943**	**0.924**	0.982	0.978	0.968
RMSEA	**0.113**	**0.139**	0.072	0.078	0.067

(b) Teacher

	Inhibit	Shift	Emotional control	Working memory	Plan /organize
$\chi 2$(d.f.) = statistic	**$\chi 2$(37) = 424.676**	**$\chi 2$(21) = 284.941**	$\chi 2$(25) = 97.434	$\chi 2$(66) = 315.156	$\chi 2$(26) = 134.791
CFI	**0.954**	**0.913**	0.988	0.968	0.976
TLI	**0.984**	**0.959**	0.996	0.992	0.988
RMSEA	**0.131**	**0.144**	0.069	0.079	0.083

CFI comparative fit index; TLI = Tucker Lewis Index; RMSEA = Root Mean Square Error of Approximation.

Table 5. Descriptive statistics and internal consistencies of the derived clinical scales by informant

(a) Parents

Clinical Scale	# items	Items	Cronbach's α
Inhibit1	4	BRIEF 3, BRIEF 33, BRIEF 38, BRIEF 54	0.769
Inhibit2	11	BRIEF 8, BRIEF 13, BRIEF 18, BRIEF 23, BRIEF 28, BRIEF 48, BRIEF 52, BRIEF 54, BRIEF 58, BRIEF 60, BRIEF 62	0.878
Shift1	5	BRIEF 5, BRIEF 15, BRIEF 30, BRIEF 35, BRIEF 45	0.767
Shift2	4	BRIEF 5, BRIEF 10, BRIEF 20, BRIEF 40	0.641
Shift3	5	BRIEF 25, BRIEF 30, BRIEF 35, BRIEF 45, BRIEF 50	0.688

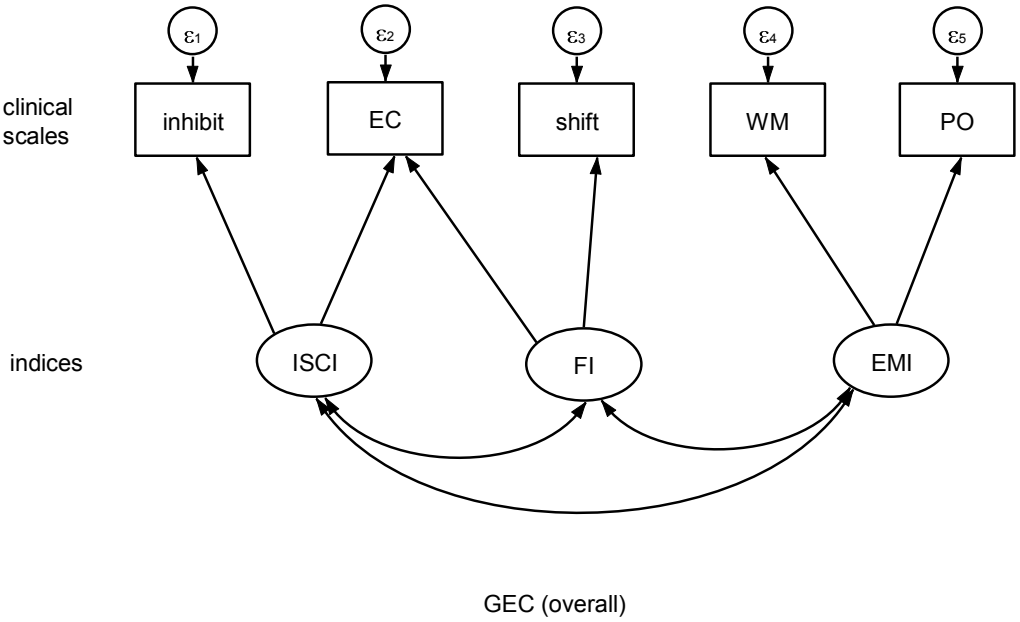
(b) Teachers

Clinical Scale	# items	Items	Cronbach's α
Inhibit1	8	BRIEF 3, BRIEF 13, BRIEF 23, BRIEF 28, BRIEF 33, BRIEF 38, BRIEF 54, BRIEF 58	0.921
Inhibit2	13	BRIEF 8, BRIEF 13, BRIEF 18, BRIEF 23, BRIEF 28, BRIEF 43, BRIEF 48, BRIEF 52, BRIEF 54, BRIEF 56, BRIEF 58, BRIEF 60, BRIEF 62	0.940
Shift1	7	BRIEF 5, BRIEF 15, BRIEF 25, BRIEF 30, BRIEF 35, BRIEF 45, BRIEF 50	0.853
Shift2	5	BRIEF 5, BRIEF 10, BRIEF 15, BRIEF 20, BRIEF 30	0.865
Shift3	4	BRIEF 10, BRIEF 20, BRIEF 30, BRIEF 40	0.795

Table 6. Test of measurement invariance by informant measurement models for the clinical scales

Model	χ^2(d.f.) = statistic	DIFFTEST χ^2(d.f.) = statistic	CFI	TLI	RMSEA
Emotional Control					
Unconstrained	$\chi^2(51)=162.198$		0.984	0.992	0.069
Constrained 1	$\chi^2(55)=248.807$	$\chi^2(12)=90.520$	0.972	0.988	0.088
Constrained 2	$\chi^2(53)=209.858$	$\chi^2(11)=64.351$	0.977	0.990	0.080
Plan/organize					
Unconstrained	$\chi^2(51)=175.863$		0.974	0.985	0.073
Constrained	$\chi^2(60)=175.764$	$\chi^2(14)=22.768$	0.976	0.988	0.065
Working memory					
Unconstrained	$\chi^2(124)=445.545$		0.962	0.989	0.075
Constrained	$\chi^2(129)=463.625$	$\chi^2(22)=79.193$	0.961	0.989	0.075

Figure 1. Measurement model of the BRIEF-P



Clinical scales: inhibit, EC (emotional control), shift, WM (working memory) and PO (plan/organize)

Indices: ISCI (Inhibitory Self-Control Index), FI (Flexibility Index) and EMI (Emergent Metacognition Index)

GEC = Global Executive Composite

Figure 2a. Measurement model for inhibit clinical scale by informant

	Inhibit I	Inhibit II
BRIEF 3	0.796	
BRIEF 8		0.376
BRIEF 13		0.646
BRIEF 18		0.759
BRIEF 23		0.713
BRIEF 28		0.558
BRIEF 33	0.906	
BRIEF 38	0.829	
BRIEF 43		0.901
BRIEF 48		0.724
BRIEF 52		0.906
BRIEF 54	0.313	0.504
BRIEF 56		0.360
BRIEF 58		0.541
BRIEF 60		0.755
BRIEF 62		0.839

parent

Inhibit I	Inhibit II
0.912	
	0.523
0.554	0.399
	0.973
0.357	0.502
0.413	0.551
0.99	
0.918	
	0.816
	0.771
	0.843
0.413	0.519
	0.432
0.415	0.378
	1.059
	0.761

teacher

Figure 2b. Measurement model for shift clinical scale by informant

	Shift I	Shift II	Shift III
BRIEF 5	0.460	0.393	
BRIEF 10		0.915	
BRIEF 15	1.072		
BRIEF 20		0.873	
BRIEF 25			0.600
BRIEF 30	0.469		0.419
BRIEF 35	0.335		0.302
BRIEF 40		0.639	
BRIEF 45	0.685		0.383
BRIEF 50			0.774

parent

Shift I	Shift II	Shift III
0.391	0.684	
	0.692	0.327
0.469	0.649	
	0.446	0.657
0.787		
0.595	0.302	0.302
0.761		
		0.876
0.676		
0.791		

teacher

Figure 3. Proposed measurement models for Executive Function - parent

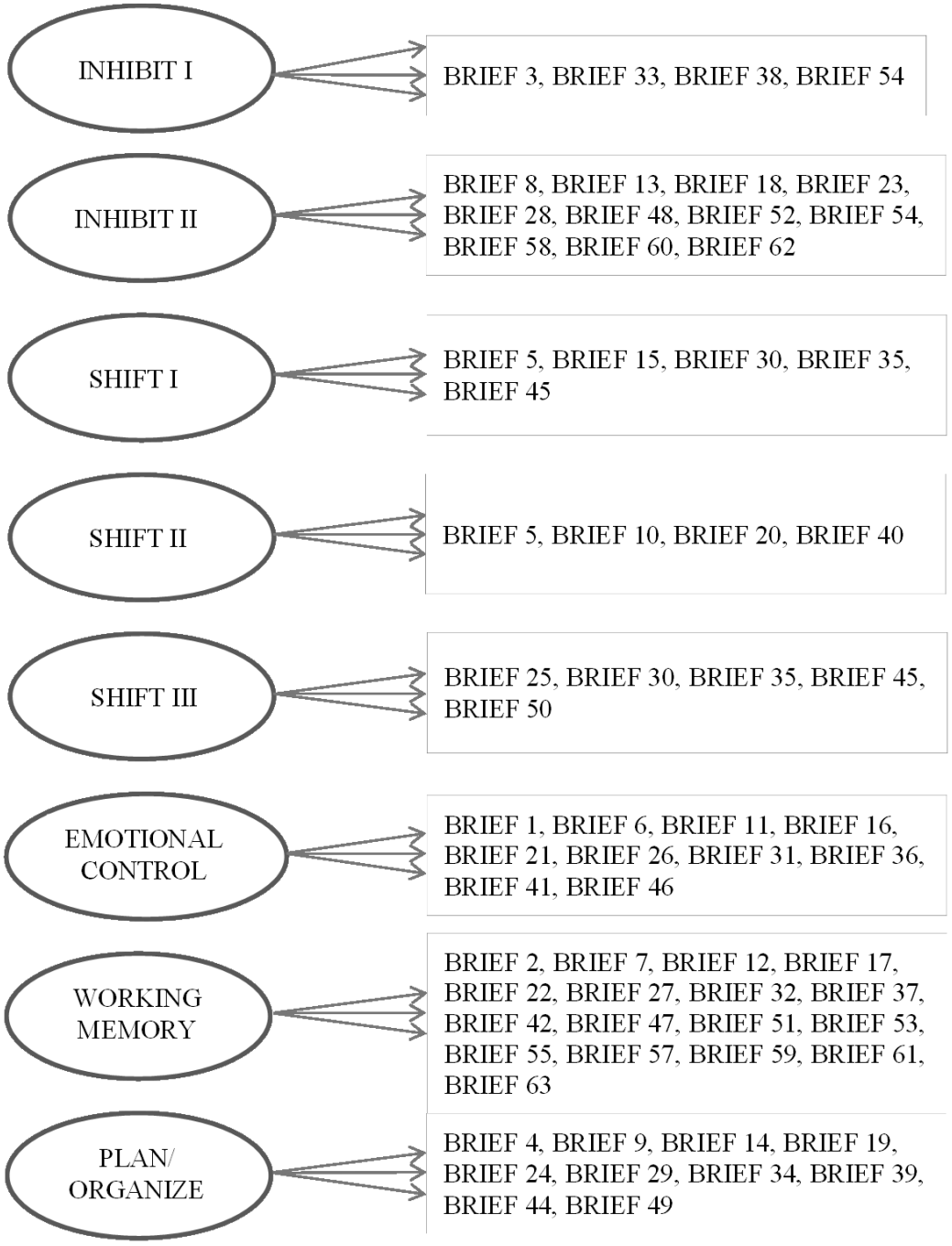
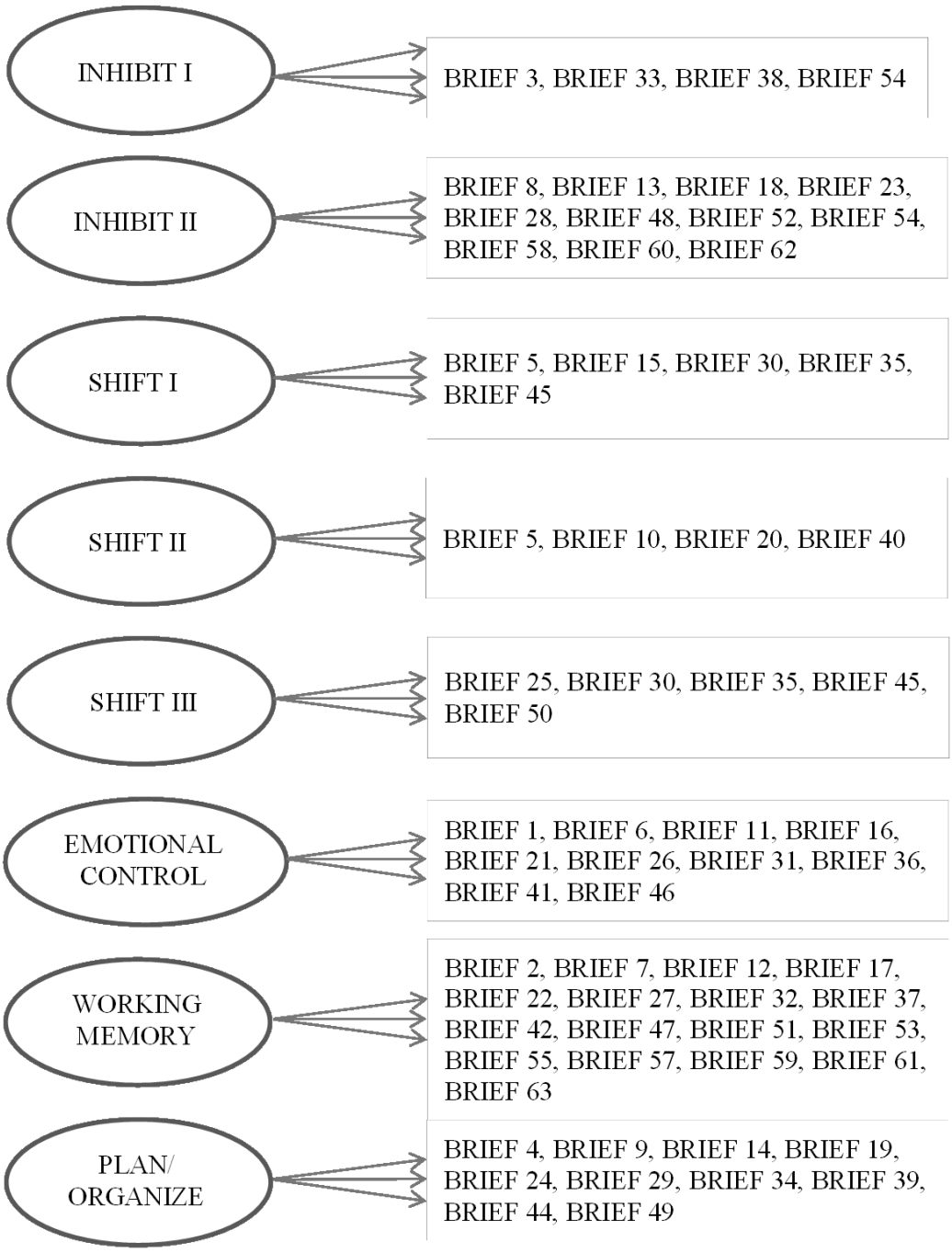


Figure 4. Proposed measurement models for Executive Function - teacher



Chapter 5. General Discussion

Measurement invariance or equivalence is important in comparative early child development (ECD) research when examining developmental differences, or changes and differences among groups, because measurement error can lead to the overestimation or underestimation of a child's score on the construct being measured (Knight & Zerr, 2010). Measurement equivalence and therefore accuracy of scientific inferences in ECD research is threatened when there are measurement problems or biases. Bias could arise because the construct may have different meanings across developmental groups or because the instrument is administered under different conditions for different groups (Byrne & Watkins, 2003). Furthermore, the credibility of inferences regarding the developmental process or developmental differences is affected by the absence of established measurement equivalence (Knight & Zerr, 2010).

The general discussion is advanced through a summary of the findings from the three studies. The research implications of the results, limitations of the overall dissertation and individual studies are discussed. This chapter ends with concluding statements of this dissertation.

Summary of Study Findings

Study 1

Study 1 is the first known comprehensive study to assess the measurement properties of the Social Responsiveness Scale (SRS; Constantino & Gruber, 2005) in a clinical sample of recently diagnosed 4-year-old preschool children with Autism Spectrum Disorder (ASD). It is also the first study to use Rasch modelling to examine the properties of the SRS in an ASD

sample. Examination of the measurement properties of the SRS in 4-year-olds is important given that many children are being referred for ASD assessment by this age (Chawarska et al., 2007) and the SRS is often used as an informative data source for clinicians (Constantino & Gruber, 2005). However, poor fit statistics and indices from the hypothesized unidimensional CCFAs showed that the 65-item SRS could not be characterized as unidimensional in our study participants. The 65-item SRS did not perform well statistically, possibly because of the age of the children. The implication of this finding is that one cannot assume measurement equivalence for any construct to be used with ASD children across as wide an age span as the SRS suggests without first testing for it. This finding also suggests that comparisons using mean construct scores between age groups and within age groups would be inappropriate without establishing the measurement invariance of the construct. Further studies using the SRS should be cautious in assuming applicability and equivalence across the age span.

Using Rasch analysis we were able to examine the measurement properties of the SRS and derive a 30-item subset of items. This 30-item subset also meets the assumptions underlying the Rasch model and therefore may have potential for use in evaluating the severity of autistic social impairment as a single dimension in other clinical samples of preschool children. The 30-item subset functioned well across age groups and could be used to represent "markers" of autistic social impairment, which would prove useful for research in preschool samples of children with autism, as it is easier to implement and yields a single-dimension construct (as proposed by Constantino & Gruber, 2005).

Study 2

Study 2 is the first known to examine the equivalence of the measurement model of the autism symptom phenotype across subgroups using algorithm items from the Autism Diagnostic

Inventory-Revised (ADI-R; Rutter et al., 2003). The results also suggest that the autism symptom phenotype is best characterized by a six factor measurement model which is stable across subgroups of ASD children. The results from this study illustrated the importance of assessing measurement equivalence prior to making comparisons between ASD subgroups. Evidence of a stable structure of the autism symptom phenotype ensures we can be more confident that the instrument is measuring the same construct across subgroups of the ASD population. Since the latent constructs used to index symptoms cannot be directly measured, having established measurement equivalence of the ADI-R, meaningful and valid comparisons across groups can be made (Borsboom, 2006). We can therefore be sure that differences in scores should reflect true differences in the construct. Furthermore, there is credibility of inferences made regarding the developmental process or developmental differences in ASD since measurement equivalence of the symptom phenotype has been established (Knight & Zerr, 2010). Stability of the autism phenotype also enables us to better monitor the progress of a child.

Study 3

Study 3, is the first known study to examine the invariance of the measurement model of executive function (EF) using the Behaviour Rating Inventory of Executive Functioning-Preschool (BRIEF-P; Gioia et al., 2003) across informants with data collected on a community sample of preschool children. It was shown that even though the BRIEF-P is a widely used instrument in research and in clinical settings, very few studies have been conducted on its measurement properties. The two studies that examined its measurement properties did so using item parcels. These studies did not formally test the equivalence of the measurement across informants (e.g., Bonillo et al., 2011; Gioia et al, 2003).

This study provided evidence that three of the clinical scales of emotional control, plan/organize and the working memory were invariant across informant. A multidimensional measurement model for EF is proposed that supports the theory that the EF construct is not unitary and is multidimensional with interrelated (or interconnected) dimensions. Further work is needed to replicate the results of this study in other studies with larger samples (in order to not run into estimation problems) and with samples in different settings (e.g., clinical). Further work is also needed on the clinical scales so that it makes use of the inevitably context-dependent nature of such ratings. The implication of these findings is that ensuring the measurement equivalence of the EF construct and clinical scales between informants is important since lack of equivalence can jeopardize cross-informant comparisons.

Research implications

An issue that needs to be investigated in the ECD literature is the implication of lack of measurement invariance on reliability and validity within groups. The extent to which lack of measurement invariance across groups translates into reduced reliability and validity is an important issue that is usually neglected in current research practices. Implications of the results of these studies for practice include the need to study groups and their differences within populations and to advocate for measurement invariance as one of the tools for constructing and validating a scale along with reliability, homogeneity and validity. These recommendations are in accordance with the guidelines published by the ITC (2010). Hence, when instruments are used with individuals from different groups (e.g., sex, education, ethnic origin, or age), then it is the responsibility of the researcher as a competent user of the instrument to make all reasonable efforts to ensure that due consideration is given to issues of fairness in testing and measurement invariance.

There are several strategies used during various stages in instrument development and use to ensure measurement invariance. One of the strategies that can be used at the design stage to establish construct equivalence across the groups of interest include the use of focus groups of experts and participants to develop the appropriate constructs (He & van de Vijver, 2012). At the pilot stage, qualitative interviews and other information can be used to determine whether or not the construct is captured. In the implementation stage, standard protocols have to be developed and adhered to by all field staff. Feedback should be collected from participants and used for further analyses and refinement of instruments. Lastly, at the analysis stage, the degree to which constructs are invariant across different groups or populations has been greatly facilitated by the development of several analytic techniques including Item Response Theory and Confirmatory Factor Analysis (Stark, Chernyshenko, & Drasgow, 2006). As researchers, it is recommended that we incorporate these strategies in all studies so as to strengthen the validity of conclusions regarding cross-group similarities and differences. Neglecting any issues of measurement invariance can lead to interpretation problems since other explanations for the differences observed cannot be ruled out.

In this dissertation and across the three studies, data collected on children during the pre-school years were obtained through parent and or teacher reports. Where ratings are obtained from different informants, and before the data are used to make any comparisons across informants, an examination of the invariance of the measurement model of the underlying construct across informant is needed. If measurement invariance cannot be established, then it is critical that we understand the reasons for differences in informant perception and differences in measurement models by informant. Even more important is the challenge of how to integrate the data from different informants because of differences in context and low agreement in ratings

between informants if invariance is established. Dirks et al. (2012) suggest that investigators examine these differences as meaningful clinical information and move toward approaches that embrace contextual variability in children's behaviour.

Measurement invariance should be a prerequisite for making any meaningful comparisons across groups since the latent constructs cannot be directly measured (Borsboom, 2006). Ideally, the differences in scores should reflect true differences in the constructs that the instruments measure. Ensuring construct comparability when testing for between-group differences, as well as differences over time is of great importance as it allows for better monitoring of health disparities in ECD across the world. Of equal importance, is establishing both comparable and equivalent constructs and items across diverse groups during early child development. In certain cases, research questionnaires and the instruments may need to be adapted when using them to study differing groups if found to be non-invariant, since the lack of equivalence can jeopardize the validity of cross-group comparisons. In addition, when researchers are developing or evaluating measurement scales, it is important to adhere to specific guidelines of item generation and to take into account the intended application of the items. A requirement of establishing measurement invariance should be included in the guidelines for comparative research studies as a necessary first step before an instrument is adopted for use and it should be a requirement for establishing validity of constructs in all ECD research.

The overarching limitation for this dissertation was in the availability of useable and accessible studies. This dissertation was therefore limited in the scope of ECD constructs that could be evaluated. Second, this dissertation was limited by the scope of data collected within each of the studies used. It was not possible to ascertain measurement invariance across

subgroups beyond those in the data collected. This is an important challenge faced in comparative research since it is not possible to develop an instrument that is invariant universally across all possible groups. Therefore, as comparisons are made beyond the original subgroups tested, it is essential the measurement invariance of the construct is examined again in different populations and subgroups as well.

Each of the studies also had some unique limitations. Limitations associated with Study 1 were that (1) the SRS was designed for use in 4- to 18-year-olds, yet our data came from preschoolers at the lower end of the age range (4 years) and (2) fewer than 100 4-year-olds were available for analyses at any single time point. We were able to use a sample of 205 4-year-olds by creating an accelerated longitudinal dataset, combining data from three time points. However it is worthy to note that age (or cohort effects) could have influenced our results.

Limitations for Study 2 included (1) the differences in the inclusion and exclusion criteria used by the different sites of the AGP consortium, (2) the sample had a high median age at administration of the ADI-R because the AGP data were collected primarily for research purposes (i.e., data were not always collected at diagnosis stage), (3) the sample was ascertained for a genetic study, there may be variation in severity of ASD symptoms compared to samples ascertained for other reasons, (4) the study was limited by the number and type of subgroups that could be evaluated to validate the phenotype[3], and (5) six ADI-R algorithm items not relevant to non-verbal children were excluded from these analyses (e.g. reciprocal conversation and social chat).

Lastly, the limitations of Study 3 included: (1) the challenge of assessing EF in preschool children based on observed performance by informant (currently there are no guidelines that

[3] For example it was not possible to examine the stability of the phenotype across IQ levels because of the different IQ measures used across the AGP data collection sites.

exist on what to do with discrepant information and how to interpret such discrepant information), (2) using data from parents and teachers as informants and the possible influence of context bias on ratings, and (3) all the children came from a daycare setting which means they had a shared experience that not all children enjoy and this may limit the generalizability of our findings to children in other care settings.

Conclusions

The pre-school period is a critical stage of development in the human lifespan (Irwin et al., 2008), and monitoring outcomes in health disparities is crucial to achieve the goal of closing the disparity, or the health gap, between nations and among subgroups within nations (CSDH, 2008). However, measurement invariance is rarely explicitly examined in ECD research and many researchers usually assume constructs are invariant across groups without checking this assumption or infer equivalence using standard psychometric tests such as test-retest reliability and content or external validity. Examination of measurement invariance, when it does occur, has been synonymous to exploring the replicability of factor structure of the construct across the groups in question (Paunonen & Ashton, 1998) even though such evidence is not sufficient on its own (Byrnes & Watkins, 2003).

Despite the progress made in measuring outcomes in early child development, researchers continue to be challenge by the groundwork needed to understand the factors that challenge validity across groups and measurement invariance and to develop instruments that minimize item bias or differential item functioning as much as possible. The results from the three studies illustrate the importance of assessing measurement invariance in ECD and whether or not the instruments examined can be used to assess subgroup differences with confidence. A lack of measurement invariance would mean that group differences in scores could be attributed,

in part, to measurement bias. More importantly, bias in the measurement of these constructs across groups can have an impact on many things including patient treatment and public policy development, particularly when there is an underlying need for a cross-group approach where belief systems may affect the meaning and structure of constructs. For example, measurement invariance of psychological tests and measurements and the valid and fair use of these tests and measurements require that these tests measure what they were intended to measure. As well, the scores from these tests should not be affected by characteristics related to membership of subgroups. Therefore, fairness in testing, as well as appropriate reporting and interpretation of results, two of the guidelines of the ITC (2010), would be undermined when measurement invariance has not been properly established for tests or measurements used to make life-altering decisions, such as treatment regimes, interventions or polices based on observed group differences.

In summary, if the assumption of measurement invariance does not hold, then differences between groups cannot be attributed to differences of the underlying construct. The source and magnitude of the non-invariance observed would need to be evaluated and decisions made as to what to do with the non-invariant items and the scale. One such recommendation would be to revise the scale or at the very least report the lack of invariance. Therefore, making improvements or changes in the practice of establishing measurement invariance of instruments used for the assessment of diverse groups of children will not only increase confidence in the outcomes of research and intervention, but can also lead to greater progress in knowledge in the various research disciplines in ECD.

References for General Introduction and Discussion

Achenbach, T. M. (1993). Implications of multiaxial empirically based assessment for behavior therapy with children. *Behavior Therapy*, *24*, 91-116.

Bandalos, D. L. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling*, *9*, 78–102.

Berry, J. W., Poortinga, Y. H., Segall, M. H., & Dasen, P. R. (2002). *Cross-cultural psychology: research and applications*. New York: Cambridge University Press.

Boltë, S., Poutska, F., & Constantino, J. N. (2008). Assessing autistic traits: cross-cultural validation of the Social Responsiveness Scale. *Autism Research, 1*, 354-363.

Bonillo, A., Araujo Jiménez, E., A., Jané Ballabriga, M., C., Capdevila, C., & Riera, R. (2012). Brief Report: Validation of Catalan Version of BRIEF-P. *Child Neuropsychology, 18*(4), 347-55.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425– 440.

Byrne, B. M., & Campbell, T. L. (1999). Cross-cultural comparisons and the presumption of equivalent measurement and theoretical structure. *Journal of Cross-Cultural Psychology*, *30*(5), pp. 555-574.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456-466.

Byrne, B. M., & Watkins, D. (2003). The issue of measurement invariance revisited. *Journal of Cross-Cultural Psychology, 34*, 155-175.

Chawarska, K., Paul, R., Klin, A., Hannigen, S., Dichtel, L. E., & Volkmar, F. (2007). Parental recognition of developmental problems in toddlers with autism spectrum disorders. *Journal of Autism and Developmental Disorders, 37*, 62-72.

Commission on Social Determinants of Health (2008). *Closing the gap in a generation: health equity through action on the social determinants of health. Final Report of the Commission on Social Determinants of Health.* Geneva: World Health Organization.

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

Dirks, M. A., De Los Reyes, A., Briggs-Gowan, M. Cella, D., & Wakschlag, L.S. (2012). Annual Research Review: Embracing not erasing contextual variability in children's behavior – theory and utility in the selection and use of methods and informants in developmental psychopathology. *Journal of Child Psychology and Psychiatry*, *53*(5): 558-574.

Duku, E., Szatmari, P., Vaillancourt, T., Georgiades, S., Thompson, A., Liu, X-Q., Paterson, A.D., & Bennett, T. (2012b). Measurement equivalence of the autism symptom phenotype in children and youth (submitted to the *Journal of Child Psychology & Psychiatry*; in press).

Duku, E., & Vaillancourt, T. (2012). Validation of the BRIEF-P in a sample of Canadian preschool children (submitted to the *Journal of Child Neuropsychology*; in press).

Duku, E., Vaillancourt, T., Szatmari, P., Georgiades, S., Zwaigenbaum, L., Smith, I. M., Bryson, S., Fombonne, E., Mirenda, P., Roberts, W., Volden, J., Waddell, C., Thompson, A., Bennett, T., & the Pathways in ASD Study Team (2012a). Investigating the Measurement Properties of the Social Responsiveness Scale in Preschool Children with Autism

Spectrum Disorders. *Journal of Autism and Developmental Disorders*. (Epub ahead of print).

Evanoff, R. (2004). Universalist, relativist, and constructivist approaches to intercultural ethics. *International Journal of International Relations,* 28, 439-458.

Gioia, G. A., Espy, K. A., & Isquith, P. K. (2003). *Behavior Rating Inventory of Executive Function, Preschool Version (BRIEF-P)*. Odessa, FL: Psychological Assessment Resources.

He, J., & van de Vijver, F. (2012). Bias and Equivalence in Cross-Cultural Research. *Online Readings in Psychology and Culture, Unit 2*. Retrieved December 22, 2012, from http://scholarworks.gvsu.edu/orpc/vol2/iss2/8.

Hoffman, E. J. (2009). Clinical features and diagnosis of autism and other pervasive developmental disorders. *Primary Psychiatry*, *16*(1), 36-44.

Horn, J. L., & McArdle, J. J. (1992). A cross-national guide to measurement invariance in aging research. *Experimental Aging Research, 18*, 117-144.

Hui, C. H., & Trandis, H. C. (1985). Measurement in cross-cultural psychology: a review and comparison of strategies. *Journal of Cross-Cultural Psychology*, *16*(2), 131-152.

International Test Commission (2001). International Guidelines for Test Use. *International Journal of Testing*, *1*(2), 93–114.

International Test Commission (2011). ITC guidelines for quality control in scoring, test analysis, and reporting of test scores. Retrieved December 21, 2012, from http://www.intestcom.org.

Irwin, L. G., Siddiqi, A. & Hertzman, C. (2008). The Equalizing Power of Early Childhood Development: From the Commission on Social Determinants of Health to Action. *Child Health and Education. 1*(3), 146-161.

Kline, R.B. (2011). *Principles and Practice of Structural Equation Modeling, 3rd edition*. New York: Guilford Press.

Knight, G. P., & Zerr, A. A. (2010). Informed theory and measurement equivalence in child development research. *Child Development Perspectives*, *4*(1), 25-30.

Kraemer, H. C., Measelle, J. R., Ablow, J. C., Essex, M. J., Boyce, W. T., & Kupfer, D. J. (2003). A new approach to integrating data from multiple informants in psychiatric assessment and research: mixing and matching contexts and perspectives. *American Journal of Psychiatry*, *160*, 1566-1577.

Lenartowicz, T., & Roth, K. (2001). Does subculture within a country matter? A cross-cultural study of motivational domains and business performance in Brazil. *Journal of International Business Studies, 32*(2), 305-325.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling*, *9*, 151–173.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13*, 127-143.

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care, 44*(11), s69-s77.

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*(4), 557-585.

Niles, F. S. (1999). Toward a cross-cultural understanding of work-related beliefs. *Human Relations*, *52*(7), 855-867.

Offord, D. R., Boyle, M. H., Racine, Y., Szatmari, P., Fleming, J. E., Sanford, M., & Lipman, E. L. (1996). Integrating assessment data from multiple informants. *Journal of the American Academy of Child and Adolescent Psychiatry*, *35*, 1078-1085.

Paunonen, S. V., & Ashton, M. C. (1998). The structured assessment of personality across cultures. *Journal of Cross-Cultural Psychology, 29*, 150–170.

Prince, M. (2008). Measurement validity in cross-cultural comparative research. *Epidemiologia e Psichiatria Sociale*, *17*(3), 211-220.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: a comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, *87*(3), 517-529.

Reichenheim, M. E., & Moraes, C. L. (2007). Operationalizing the cross-cultural adaptation of epidemiological measurement instruments. *Rev Saúde Pública, 41*(4), 665-73.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological Bulletin, 114*, 552-566.

Rottig, D. (2009). Overcoming common pitfalls in cross cultural management research. *International Business: Research Teaching and Practice, 3*(1), 32-51.

Rutter, M., Le Couteur, A., & Lord, C. (2003). *Autism Diagnostic Interview-Revised*. Los Angeles, USA: Western Psychological Services.

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: review of practice and implications. *Human Resource Management Review*, *18*, 210–222.

Stark, S., Chernyshenko, S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292-1306.

Steenkamp, J. E. M., & Baumgartner, H. (1998), "Assessing Measurement Invariance in Cross-National Consumer Research," *Journal of Consumer Research*, *25*, 78-90.

Steinmetz, H., Schmidt, P., Tina-Booh, A., Wieczorek, S., & Schwartz, S. H. (2007). Testing measurement invariance using multigroup CFA: differences between educational groups in human values measurement. *Quality & Quantity*. doi 10.1007/s11135-007-9143-x

Szatmari, P., Liu, X. Q., Goldberg, J., Zwaigenbaum, L., Paterson, A. D., Woodbury-Smith, M., Georgiades, S., Duku, E., & Thompson, A. (2012). Sex differences in repetitive stereotyped behaviors in autism: Implications for genetic liability. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *159B*(1), 5–12.

Taris, T. W., Bok, I. A., & Meijer, Z. Y. (1998). Assessing stability and change of psychometric properties of multi-item concepts across different situations: a general approach. *Journal of Psychology, 132*(3), 301-316.

Teresi, J. A. (2006). Overview of quantitative measurement methods equivalence, invariance, and differential item functioning in health applications. *Medical Care*, *44*(11), s39-s49.

Van de Vijver, F. J. R., & Leung, K. (1997). Methods and data analysis of comparative research. In J. W. Berry, Y. H. Poortinga, & J. Pandey (Eds.), *Handbook of cross-cultural psychology* (2nd ed.) (pp. 257-300). Boston: Allyn & Bacon.

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods*, *5*(2), 139-158. Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement

invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4-69.

Weston, R., & Gore Jr, P. A. (2006). A brief guide to structural equation modeling. *The Counseling Psychologist*, *34*(5), 719-751.