From the Department of Women's and Children's Health
Karolinska Institutet, Stockholm, Sweden

STANDARDIZING DIAGNOSTIC ASSESSMENT OF AUTISM SPECTRUM DISORDER - ASSETS AND CHALLENGES

Eric Zander

Karolinska Institutet

Stockholm 2015

Standardizing diagnostic assessment of Autism Spectrum Disorder - assets and challenges

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Eric Zander

Principal Supervisor:
Professor Sven Bölte
Karolinska Institutet
Department of Women's and Children's Health
Neuropsychiatry Unit
Center of Neurodevelopmental Disorders (KIND)

Co-supervisor:
Associate Professor Terje Falck-Ytter
Karolinska Institutet
Department of Women's and Children's Health
Neuropsychiatry Unit
Center of Neurodevelopmental Disorders (KIND)

Uppsala University
Department of Psychology
Uppsala Child and Baby Lab

Opponent:
Professor Mikael Heimann
Linköping University
Department of Behavioural Sciences and Learning

Swedish Institute of Disability Research
Linköping University

Examination Board:
Associate Professor Eva Billstedt
University of Gothenburg
Sahlgrenska Academy
Institute of Neuroscience and Physiology
Gillberg Neuropsychiatry Centre

Professor Ulla Ek
Stockholm University
Department of Special Education

Associate Professor Predrag Petrovic
Karolinska Institutet
Department of Clinical Neuroscience

To Ulla

ABSTRACT

The overall aim of this thesis was to study some previously under-researched psychometric properties of the standardized diagnostic instruments Autism Diagnostic Interview-Revised [ADI-R] and Autism Diagnostic Observation Schedule [ADOS] as well as the standardized Vineland Adaptive Behavior Scales, Second Edition [VABS-II] in order to improve current diagnostic practice.

Excellent interpersonal objectivity operationalized as interrater reliability of the ADI-R and the ADOS has been reported from research settings with highly trained, research reliable examiners. However, there are no studies from clinical practice despite the wide usage of both instruments in these settings. In studies I and II, the objectivity of the ADI-R and the ADOS was examined in clinical practice among clinical users. For the ADI-R, the objectivity for items (medians), domains and criteria exceeded $G(q,k)$ (analogous to intraclass correlation) $= .90$ in all instances and was $\kappa = .83$ for classification. For modules 1-4 of the ADOS, items (medians) ranged from $G(q,k) = .74$ to $.83$, overall totals from $.85$ to $.92$ and classification was $\kappa = .69$.

Diagnostic validity of the ADI-R's and the ADOS' revised algorithms, separately and in combination, has previously only occasionally been reported for young children. In study III, the diagnostic validity of these instruments was examined in a large clinical sample of toddlers and young preschoolers. Diagnostic validity for the combined ADI-R and the ADOS yielded a sensitivity of 78% and specificity of 88% while the classification accuracy for the single use of the ADI-R and the ADOS in general was lower.

Finally, despite a multitude of studies reporting on the effect of the new DSM-5 symptomatology criteria of ASD on diagnosis rates, no studies have analyzed the effect of the novel DSM-5 impairment criterion. In study IV, this was examined in a subsample of the young children of study III. In accordance with the DSM-5 conceptualization of impairment, this was measured with the standardized VABS-II and different cutoffs of adaptive functioning were applied to operationalize impairment. We showed that, depending on the threshold (1, 1 ½ and 2 SDs below the mean) used to operationalize impairment, 88%, 69% and 33% respectively of the children with a DSM-IV-TR ASD diagnosis fulfilled the impairment criterion compared to 91% for the DSM-5 symptomatology criteria.

To conclude, in terms of objectivity our results endorse the universal use of the ADI-R and the ADOS in everyday clinical practice and research for all age groups as the first choice diagnostic instruments for ASD. The combined use of the ADI-R and the ADOS yields better diagnostic validity than the separate usage of either instrument. Moreover, our results suggest that a strict application of the new DSM-5 impairment criterion might compromise early diagnosis for 12-67% of young children with a complete DSM-5 ASD symptomatology. The main asset of using standardized instruments is improved objectivity in assessing ASD symptomatology.

SCIENTIFIC PAPERS INCLUDED IN THE THESIS

I. **Eric Zander**, Charlotte Willfors, Steve Berggren, Christina Coco, Anette Holm, Ida Jifält, Renata Kosieradzki, Jenny Linder, Viviann Nordin, Karin Olafsdottir, & Sven Bölte. (2015). The objectivity of the Autism Diagnostic Interview-Revised (ADI-R) in clinical settings. Submitted for publication.

II. **Eric Zander**, Charlotte Willfors, Steve Berggren, Nora Choque-Olsson, Christina Coco, Anna Elmund, Åsa Hedfors Moretti, Anette Holm, Ida Jifält, Renata Kosieradzki, Jenny Linder, Viviann Nordin, Karin Olafsdottir, Lina Poltrago, & Sven Bölte. (2015). The objectivity of the Autism Diagnostic Observation Schedule (ADOS) in naturalistic clinical settings. *European Child & Adolescent Psychiatry*, 2015 Nov [Epub ahead of print]. doi:10.1007/s00787-015-0793-2.

III. **Eric Zander**, Harald Sturm, & Sven Bölte. (2015). The added value of the combined use of the Autism Diagnostic Interview-Revised and the Autism Diagnostic Observation Schedule: Diagnostic validity in a clinical Swedish sample of toddlers and young preschoolers. *Autism*, 19(2), 187-199. doi:10.1177/1362361313516199.

IV. **Eric Zander**, & Sven Bölte. (2015). The New DSM-5 Impairment Criterion: A Challenge to Early Autism Spectrum Disorder Diagnosis? *Journal of Autism and Developmental Disorders*, 45(11), 3634-3643. doi:10.1007/s10803-015-2512-8.

SCIENTIFIC PAPERS NOT INCLUDED IN THE THESIS

I. **Eric Zander**, & SvenOlof Dahlgren. (2010). WISC-III index score profiles of 520 Swedish children with pervasive developmental disorders. *Psychological assessment*, 22(2), 213-222. doi:10.1037/a0018335.

II. Selma Idring, Dheeraj Rai, Henrik Dal, Christina Dalman, Harald Sturm, **Eric Zander**, Brian K. Lee, Eva Serlachius, & Cecilia Magnusson. (2012). Autism spectrum disorders in the Stockholm Youth Cohort: design, prevalence and validity. *PLoS One*, 7(7), e41280. doi:10.1371/journal.pone.0041280.

III. Jacob Åsberg, Ulla Zander, **Eric Zander**, & Annika Dahlgren Sandberg. (2012). Social and Individual Aspects of Classroom Learning in Students with Autism Spectrum Disorder: An Action Research Pilot Study on Assessment. *European Journal of Special Needs Education*, 27(1), 115-127. doi:10.1080/08856257.2011.640488.

IV. Annelies de Bildt, Sjoerd Sytema, **Eric Zander**, Sven Bölte, Harald Sturm, Nurit Yirmiya, Maya Yaari, Tony Charman, Erica Salomone, Ann LeCouteur, Jonathan Green, Ricardo Canal Bedia, Patricia García Primo, Emma van Daalen, Maretha V. de Jonge, Emilía Guðmundsdóttir, Sigurrós Jóhannsdóttir, Marija Raleva, Meri Boskovska, Bernadette Rogé Sophie Baduel, Irma Moilanen, Anneli Yliherva, Jan Buitelaar, & Iris J. Oosterling. (2015). Autism Diagnostic Interview-Revised (ADI-R) Algorithms for Toddlers and Young Preschoolers: Application in a Non-US Sample of 1,104 Children. *Journal of Autism and Developmental Disorders*, 45(7), 1-16. doi:10.1007/s10803-015-2372-2.

V. Pär Nyström, Gustaf Gredebäck, Sven Bölte, Terje Falck-Ytter, & the EASE team*. (2015). Hypersensitive pupillary light reflex in infants at risk for autism. *Molecular Autism,* 6, 10. doi:10.1186/s13229-015-0011-6.

* the EASE team: Linda E. Campbell, Elodie Cauvet, Therese L Ekberg, Johan Lundin Kleberg, Elisabeth Nilsson Jobs, Emilia Thorup, **Eric Zander**

CONTENTS

LIST OF ABBREVIATIONS

3di	Developmental, Dimensional and Diagnostic Interview
ABAS-2	Adaptive Behavior Assessment System, Second Edition
ADHD	Attention Deficit/Hyperactivity Disorder
ADI-R	Autism Diagnostic Interview-Revised
ADOS/-2	Autism Diagnostic Observation Schedule/-Second Edition
APA	American Psychiatric Association
ASD	Autism Spectrum Disorder
ASDI	Asperger Syndrome (and High-Functioning Autism) Diagnostic Interview
AUC	Area under the curve
C-GAS	Children's Global Assessment Scale
CDC	Center of Disease Control
CI	Confidence Interval
CIDI	Composite International Diagnostic Interview
COM	Qualitative abnormalities in communication
CS/ CSS	Comparison Score/ Calibrated Severity Scores (used synonymously)
d	Cohen's estimat for effect size
DAWBA	Development and Well-Being Assessment
DISCO	Diagnostic Interview for Social and Communication Disorders
DSM	Diagnostic and statistical manual of mental disorders
DSM-IV-TR	Diagnostic and statistical manual of mental disorders-Fourth Edition-Text Revision
FN	False Negatives: children with ASD falsely classified as NS
FP	False Positives: NS children falsely classified as ASD
G(q,k)	G = G-Theory; q = mean number of raters that each ratee share in common, i.e. a multiplier that scales the contribution of variance attributable to rater main effects; k =number of raters
GAF	Global Assessment of Functioning
ICC	Intraclass Correlation

ICD-10	International Classification of Diseases-Tenth Edition
ICF/-CY	International Classification of Functioning, Disability and Health/-Version for Children and Youth
ID	Intellectual Disability
IGP	Imitation, Gestures, and Play
IQ	Intelligence Quotient
ISMD	Ill-Structured Measurement Design
K-SADS-PL	Kiddie Schedule for Affective Disorders and Schizophrenia for School Aged Children - Present and Lifetime version
LEAD	Longitudinal, Expert, and All Data
LR+	Positive likelihood ratio: sensitivity/(1–specificity)
MRI	Magnet Resonance Imaging
NS	Non Spectrum (not ASD)
NV / 12-20/NV21-47	ADI-R developmental cell: all children aged 12–20 months + nonverbal children aged 21–47 months
NVIQ	Non Verbal Intelligence Quotient
OR	Odds Ratio
PA	Percent Exact Agreement
PDD-NOS	Pervasive Developmental Disorder Not Otherwise Specified
PEP-3	Psychoeducational Profile – Third Edition
PH / PH21-47	ADI-R developmental cell: children aged 21–47 months with phrase speech
PIA	Parent Interview for Autism
κ	Kappa
κ_w	Weighted kappa
r	Pearson Product Correlation Coefficient
RDoC	Research Domain Criteria
ROC	Receiver Operating Characteristics
RPI	Reciprocal Peer Interaction
RRB	Repetitive and Restricted Behavior
RSI	Qualitative Abnormalities in Reciprocal Social Interaction
SA	Social Affect

SBU	Swedish Council on Health Technology Assessment
SC	Social Communication
SCID	Structured Clinical Interview for DSM-IV
SD	Standard Deviation
SEK	Swedish Krona
SRS	Social Responsiveness Scale
SW / SW21-47	ADI-R developmental cell: children aged 21–47 months with single words
TN	True Negatives: NS children correctly classified as NS
TP	True Positives: children with ASD correctly classified as ASD
UK	United Kingdom
US	United States of America
VABC	Vineland-II Adaptive Behavior Composite
VABS/-II	Vineland Adaptive Behavior Scale/-Second Edition
VIQ	Verbal IQ
WHO	World Health Organization
WHODAS	World Health Organization Disability Assessment Schedule
χ^2	Chi-square statistics

1 BACKGROUND

1.1 DEFINITION AND DESCRIPTION OF AUTISM SPECTRUM DISORDER

Autism Spectrum Disorder (ASD) is a group of early onset neurodevelopmental disorders that are behaviorally defined by a combination of persistent deficits in social communication and social interaction alongside stereotypic, restricted and repetitive behaviors causing functional impairment (American Psychiatric Association [APA], 2000, 2013; World Health Organisation [WHO], 1993). Evidence indicates neurobiological underpinnings for ASD causing altered brain functions related to specific cognitive dysfunctions that in turn could be linked to the defining behavioral pattern (Lai, Lombardo, & Baron-Cohen, 2013). Currently, the idea of atypical brain connectivity has gained great interest as a main characteristic in ASD (Rane et al., 2015). In that context and in general, there is an intense ongoing research effort on the volumetric, structural and functional aspects of the ASD brain with different imaging techniques (Philip et al., 2012) and post-mortem studies (Casanova, 2014) as well as in the field of biochemistry (Lam, Aman, & Arnold, 2006) and immunology (Hsiao, 2013; Noriega & Savelkoul, 2014). Research has shown that genetics have a strong impact on ASD. Estimates of heritability from twin studies have ranged from 38% (Hallmayer et al., 2011; Lichtenstein, Carlström, Råstam, Gillberg, & Anckarsäter, 2010) to over 90% (Bailey et al., 1995; Colvert et al., 2015), but a recent well-designed study with a very large sample reported a heritability estimate of 50% and a dose-response relationship of individual risk of ASD proportionally linked to how closely related individuals were with each other (Sandin et al., 2014). However, the genetic landscape of ASD is characterized by an extreme complexity and remarkable heterogeneity where a convergent pattern of pathways and mechanisms affected by the currently known risk genes just has begun to emerge (Geschwind & State, 2015). To date, genetic antecedents have only been identified in a minority of individuals with ASD, and these are often characterized by pleiotropy and inherited as well as rare and common de novo mechanisms seem to be at play. There are a range of known and potential environmental risk factors associated to ASD influencing the development, already and especially before conception and birth, but neither is necessary nor sufficient to develop ASD (Lyall, Schmidt, & Hertz-Picciotto, 2014). Etiologically speaking, the dominating view is still that ASD results of a complex interaction between a large part genetic but also epigenetic, environmental and developmental factors (Lai et al., 2013).

Common cognitive theories of ASD that have also been linked to differences in brain functioning compared to healthy controls are deficits in the ability to mentalize (Theory of Mind), i.e. the intuitive understanding of mental states in self and others (Boucher, 2012; Pelphrey, Shultz, Hudac, & Vander Wyk, 2011; Senju, 2012), executive dysfunctions (Leung & Zakzanis, 2014) and weak central coherence, i.e. a cognitive style prioritizing details over the global picture (Happé & Frith, 2006; Samson, Mottron, Soulières, & Zeffiro, 2012).

1.2 PREVALENCE

Current prevalence estimates of ASD indicate that around 1 % of the population have ASD (CDC, 2014; Elsabbagh et al., 2012) but some recent register data from Stockholm and prevalence estimates from South Korea report rates of ASD in the range of 2.5% (Idring et al., 2015; Y. S. Kim et al., 2011). There has been a huge and steady increase since the first prevalence study in 1966 reported an estimate of less than 0.5 ‰ (Lotter, 1966) commonly explained by changes of the definition of ASD and in reporting practices (Hansen, Schendel, & Parner, 2015) as well as increased awareness (Elsabbagh et al., 2012) even though an increase due to risk factors cannot be ruled out (Lai et al., 2013). Males are 2-4 times more prevalent than females and concurrent intellectual disability [ID], language disorders as well as other neurodevelopmental, psychiatric and neurological disorders are common in ASD (Joshi et al., 2011; Simonoff et al., 2013).

1.3 CONSEQUENCES AND INTERVENTIONS

Having ASD is from early on associated with personal suffering and considerable challenges for families. Without a diagnosis and in a non-autism friendly environment, the social communication difficulties accompanied by inflexible and sometimes odd thought and behavioral patterns often result in frequent conflicts, peer rejection (Carter, Garrity-Rokous, Chazan-Cohen, Little, & Briggs-Gowan, 2001; Williamson, Craig, & Slinger, 2008), bullying (Humphrey & Symes, 2011; Schroeder, Cappadocia, Bebko, Pepler, & Weiss, 2014; van Roekel, Scholte, & Didden, 2010; Zablotsky, Bradshaw, Anderson, & Law, 2014), loneliness (Bauminger, Shulman, & Agam, 2003; Lasgaard, Nielsen, Eriksen, & Goossens, 2010; White & Roberson-Nay, 2009), and general adaption problems affecting mental health, learning, quality of life and social inclusion negatively (Barneveld, Swaab, Fagel, van Engeland, & de Sonneville, 2014; Domellöf, Hedlund, & Ödman, 2014; Kuhlthau et al., 2010; van Heijst & Geurts, 2015). There exists still no pharmacological or other cure for ASD but a wide range of interventions and services are available for individuals with ASD and their families (Bölte, 2014). The main focus of these interventions is on developing (deficient) skills, reducing unwanted behaviors, and implementing autism friendly environments facilitating communication and daily functioning in preschool, school and at work as well as at home. Thus, many interventions are educational activities for parents, teachers and other persons dealing with the individual, as well as communication strategies, behavior modification and positive behavior support directly addressing the individual affected by ASD (Hirvikoski et al., 2015). Adult outcome in ASD is mixed but often poor with a need for continuous support throughout the life span for many individuals (Billstedt & Gillberg, 2005; Henninger & Taylor, 2013; Magiati, Tay, & Howlin, 2014; L. E. Smith, Greenberg, & Mailick, 2012). Early intervention is assumed to improve outcome (Fein et al., 2013; Reichow, 2012), which is one reason for promoting early detection and early diagnosis around the world. In many cases it is possible to assign a reliable ASD diagnosis around the age of two years but the average age of diagnosis is still higher (CDC, 2014; Idring et al., 2012). Research shows that being a parent of an individual with ASD is associated with more stress (Hayes & Watson,

2013), lower incomes (Jacob, Scott, Falkmer, & Falkmer, 2015) and higher rates of sick leave and unemployment (McEvilly, Wicks, & Dalman, 2015; Ou et al., 2015). The lifetime costs of ASD have been estimated to between 10 and 15 million SEK per individual (Buescher, Cidav, Knapp, & Mandell, 2014; Gustavsson et al., 2011; Järbrink, 2007; Knapp, Romeo, & Beecham, 2009; Lavelle et al., 2014; Leigh & Du, 2015). In summary, ASD constitutes a substantial public health issue affecting the life of many individuals, not only those directly affected with ASD.

1.4   DIAGNOSING AUTISM SPECTRUM DISORDER

Assigning reliable and valid diagnoses of ASD and other mental disorders lies at the heart of clinical practice and research in psychiatry. Clinically, a correct diagnosis is a necessary basis to understand the behavior and needs of the individual, plan intervention and give access to service. In psychiatry research, diagnosis serves as the launch pad and heuristically organizational principle for studying for example epidemiology, etiology, neurobiology, genomics, cognition, behavior, treatment responses, trajectories and outcome.

To diagnose means to determine to which diagnostic category or categories an individual belong according to the current diagnostic classifications. The psychiatric nosology is based on the conceptualization of mental disorders as disturbances of mental functioning reflecting underlying dysfunctions in biological, developmental or psychological processes (APA, 2013). As we still lack complete knowledge of these latent, underlying processes, we rely on manifest patterns of mostly behavioral symptoms, summarized by the diagnostic criteria of the published classifications, to define mental disorders (Hofmann, 2014). Therefore, even though these sets of diagnostic criteria are seen as the best available description of mental disorders to date, it is also explicitly acknowledged that they constitute but an incomplete and not fully validated definition of the disorder (APA, 2013; WHO, 1993; 1992). In fact, research so far has not been successful in convincingly linking many of the disorders of the current classifications to common genetic, neurobiological and brain functioning patterns why a new research paradigm, the Research Domain Criteria (RDoC), seeking to organize psychiatric knowledge around pathophysiological entities instead of the current behaviorally defined diagnoses, has been established but not yet affected the classifications like for instance the DSM (Insel et al., 2010; London, 2014). So while waiting for a new RDoC based classification of mental disorders, diagnostic criteria of the current classifications, then, are offered as "guidelines" to be applied with clinical judgment by specifically trained and experienced clinicians to make reliable and valid diagnoses (APA, 2013, p.21). Moreover, a necessary but not sufficient part of the process of assigning accurate psychiatric diagnoses consists of assessing symptom presentations in relation to the published criteria as reliably as possible. Furthermore, an evaluation of intellectual and language functioning, genetic and environmental risk factors, comorbid developmental, psychiatric or neurological conditions, the course of the difficulties, etc. are also warranted to accurately assign valid diagnoses.

1.5 AUTISM SPECTRUM DISORDER ACCORDING TO THE CURRENT MEDICAL CLASSIFICATIONS OF MENTAL DISORDERS

1.5.1 The behavioral psychopathological criteria

In both WHO's International Classification of Diseases [ICD-10] and APA's Diagnostic and Statistical Manual of Mental Disorders [DSM-IV/-TR and DSM-5], the behavioral symptomatology of ASD includes persistent deficits in social communication and social interaction together with patterns of restricted, repetitive and stereotyped behaviors, interests, and activities (RRB). In ICD-10 and DSM-IV/-TR, social interaction and communication deficits are divided into two separate criteria with 4 items each while DSM-5 combines these two domains into one single criterion with 3 sub-criteria, see Appendix for complete diagnostic criteria of the three classifications. The RRB criterion is subdivided in 4 symptoms/items in ICD-10, DSM-IV/-TR and DSM-5 but with some differences in content. However, despite these differences in structure across diagnostic systems, the totality of the psychopathological content covered is mainly the same. ICD-10 and DSM-IV/-TR share the same structural organization of criteria and with only minor differences in wording and content. DSM-5 combines all verbal and non-verbal communication aspects into two sub-criteria (B1-2), converts the language delay/absence part of the ICD-10 B2a and DSM-IV/-TR A2a criterion to a specifier and moves the stereotyped language part to the RRB section. In the RRB section, sensory behaviors and reactions to sensory stimuli constitute a new sub-criterion while several of the former ICD-10 and DSM-IV/-TR items/symptoms have been somewhat expanded and reorganized. According to ICD-10 and DSM-IV/-TR, it requires 6 symptoms/items with the following distribution: at least two on social interaction and at least one on communication and RRB respectively. DSM-5 requires that all three parts of social communication/interaction and at least two of the four RRB parts are or have been present (current or by history).

1.5.2 The onset criterion and exclusion criteria

All three classifications require early onset (prior to age three years or in the early developmental period, ICD-10: A; DSM-IV/-TR: B; DSM-5: C). Symptoms must not be better explained by general developmental delay (low IQ) in ICD-10 (C) and DSM-5 (E), and/or by Rett's Disorder in DSM-IV/-TR (C) and ICD-10 (C) or Childhood Disintegrative Disorder in DSM-IV/-TR (C) as well as by any other developmental disorders or psychiatric diagnoses in ICD-10.

1.5.3 The impairment criterion

Moreover, DSM-5 includes a new criterion: D. "Symptoms cause clinically significant impairment in social, occupational, or other important areas of current functioning" (APA, 2013, p.50), only required for Asperger's disorder in DSM-IV/-TR and not at all in ICD-10. The new DSM-5 impairment criterion for ASD reflects a change in the conceptualization of mental disorders compared to earlier version of DSM. Since at least DSM-III (APA, 1980),

functional impairment (or distress) has been a completing criterion for many diagnoses, and is still an additional criterion in DSM-5 for many diagnoses including ASD. From having basically served as a threshold criterion, and conceptually intermingled with symptomatology, DSM-5 promotes a new view of functional impairment as an important dimension separated from symptomatology in the assessment and intervention of mental disorder. This harmonizes the DSM with the system of the WHO where symptomatology and functioning are separate constructs and treated in different classifications, ICD-10 (WHO, 1993) and the International Classification of Functioning, Disability and Health for adults as well as for children and youths [ICF/ICF-CY] (WHO, 2001; 2007). Moreover, according to the DSM-5 conceptualization of mental disorder, impairment is caused by the latent disease entity and its symptoms. There is no explicit information in DSM-5, though, of how impairment in children and youths should be defined, operationalized or measured although the use of the World Health Organization Disability Assessment Schedule (WHODAS, Üstün et al., 2010), based on the ICF, is recommended for adults. It has been demonstrated that there is a close relationship between the concept of the ICF components activities and participation, that are the basic components of the WHODAS, and adaptive functioning (Fabiano & Pelham Jr., 2009). Impairment is then the result of "activity limitation" and "participation restrictions" according to the ICF or limited functioning according to the adaptive functioning concept.

1.5.4 Diagnostic subcategories

In DSM-5, autistic and Asperger's disorder and Pervasive Developmental Disorder Not Otherwise Specified [PDD-NOS] have merged into one diagnostic category, Autism Spectrum Disorder; Rett's Disorder and Childhood Disintegrative Disorder have been omitted from DSM-5, the former because it has an established neurological basis where ASD symptoms may be present for a period and the latter because there is insufficient scientific evidence of its existence as a separate category.

1.5.5 Specifiers

In DSM-5, a set of specifiers (not criteria) is introduced for the ASD diagnosis. Symptom severity for the social communication and social interaction difficulties and the RRB should be specified separately as well as the presence (or absence) of intellectual and/or language impairment, known medical, genetic, or environmental factors, other neurodevelopmental, mental or behavioral disorders and catatonia. It is acknowledged that these factors might influence the way the symptomatology is expressed in the individual.

1.5.5.1 Severity

The current diagnostic classificatory systems are criterion-based and categorical, meaning that clinicians have to assess if a person fulfills the required criteria as well as belongs to diagnostic category or not. However, there is also a dimensional approach present for many diagnoses including ASD as the evaluation and documentation of disorder severity now is

mandatory. ASD severity, like the presence or absence of intellectual and language impairment, and neurological, (neuro)developmental or psychiatric comorbidities, is a specifier and linked to how much support an individual needs because of the functional impairment his or her ASD symptoms cause. In general within the framework of the current diagnostic classification systems, severity has been operationalized as symptom count and/or an evaluation of the intensity of each symptom (Bernier, 2012). For ASD according to DSM-5, the relationship between the severity specifier and the impairment criterion is not specified.

1.5.6 DSM-IV/-TR versus DSM-5

Since the first drafts of the new ASD criteria of DSM-5 were made public, there has been an ongoing discussion and even controversy of how the new symptomatology criteria will influence diagnosis rates. A multitude of studies have concluded that the application of the new criteria would reduce diagnostic rates with at least third compared to DSM-IV/-TR and even more for the higher functioning individuals with ASD (summarized in Kulage, Smaldone, & Cohn, 2014; I. Smith, Reichow, & Volkmar, 2015; Sturmey & Dalfern, 2014; see also Barton, Robins, Jashar, Brennan, & Fein, 2013; Turygin, Matson, Adams, & Belva, 2013; Volkmar & Reichow, 2013). However, to date, there are very few studies analyzing the potential impact of the application of the new impairment criterion on diagnosis rate or with a few exceptions even mentioning it (Maenner et al., 2014; Weitlauf, Gotham, Vehorn, & Warren, 2014).

1.5.7 Factors influencing the manifestation of the behavioral symptomatology

For ASD as for neurodevelopmental disorders in general, at least since Michael Rutter formulated the modern behavioral definition of autism in a seminal paper in the seventies, drawing on Leo Kanner's original description of the clinical phenomena he named autism (Kanner, 1943) and research, the necessity to take certain unspecific factors into account when assessing ASD symptomatology have been evident (Rutter, 1978). Rutter highlighted the importance of putting the ASD symptomatology in relation to age and IQ/mental age to correctly characterize it, because of the observation that the absence or immaturity of certain behavioral skills in very young children or children with a very low mental age mistakenly could be interpreted as autistic symptoms as well as the often marked change and improvement of behaviors in individuals with autism above 5 years of age. The influence of age and/or developmental level and expressive language level on the symptomatology have been corroborated in more recent research (Gotham, Pickles, & Lord, 2009; Hus & Lord, 2013) and both intellectual and language impairment are specifiers for an ASD diagnosis since DSM-5 because of that. Therefore, to accurately assess the ASD symptomatology, a detailed evaluation of the individual's intellectual and language functioning is necessary.

1.6 DIAGNOSTIC EVALUATION OF AUTISM SPECTRUM DISORDER

For the case formulation process or diagnostic evaluation of ASD and the other neurodevelopmental disorders, a multidisciplinary approach is often recommended, traditionally centered around a detailed clinical history but also relying on other methodologies to adequately characterize the individual's psychopathology and need for treatment and service (NICE Clinical Guidelines, 2011; Swedish Council on Health Technology Assessment [SBU], 2013; Volkmar, Langford Booth, McPartland, & Wiesner, 2014; Volkmar, Siegel, et al., 2014). Apart from clinical expertise, it is generally known that the use of standardized diagnostic instruments, i.e. checklists, interviews and observational methods, with good psychometric properties could improve the reliability of diagnosis (Wittchen, 1994), and in DSM-5 (APA, 2013) this is also explicitly endorsed for the evaluation of ASD (p. 55).

1.6.1 Standardized diagnostic instruments

Standardized diagnostic instruments operationalizing diagnostic criteria aim to reduce variance in how clinicians collect, process and interpret information through the use of a shared set of procedures. For example, checklists and structured and respondent-based interviews comprise mandatory verbatim questions and response alternatives from which the respondent choses what to answer, whereas in an investigator-based structured interview, the interviewer asks mandatory verbatim but often open-ended questions, follow rules for probing and independently assesses and codes the respondent's answers according to some kind of coding scheme. Standardized observation instruments have mandatory administration rules, specified activities and standard material to use as well as rules of what to quantify and how. Most standardized instruments also include algorithms, i.e. set of rules of how to calculate the overall results and then how to interpret these results in terms of fulfilling diagnostic criteria or not, for example against a cutoff or according to different kinds of norms. Specific training is generally prescribed for using standardized diagnostic instruments. Thus, the structured and consistent way of collecting and evaluating information in relation to diagnostic criteria means that all relevant areas will be covered and that information variance, observation/interpretation variance, and criterion variance are minimized (Spitzer, 1983; Wittchen, 1994).

1.6.2 Reliability, objectivity and validity of standardized diagnostic instruments

The reliability and the validity of standardized instruments represent quantified information of the measurement precision, consistency and repeatability of the instrument, i.e. how well it discriminates subjects from each other, as well as its conformity with truth, i.e. how well it measures what it is intended to measure. Reliability is basically empirical and its most common forms in regards to diagnostic instruments are internal consistency, a measure of the agreement among the items of the instrument and test-retest reliability, which deals with the temporal stability of the result. The interrater reliability, sometimes only reliability (APA,

2013; Freedman et al., 2013), that measures the agreement between different observers or more correctly the extent to which the obtained result is independent of the examiner is often treated as an aspect of reliability (Blacker and Endicott, 2000) but represents according to classical test theory the (interpersonal) objectivity aspect of an instrument (Lienert, 1989). Objectivity, used interchangeably with the term interrater reliability in this thesis, is a crucial property of every standardized diagnostic instrument that involves the coding of observed behaviors or verbal descriptions of behaviors. Although the objectivity of a measure generally is seen as property of the instrument, it is also dependent on external factors such as sample characteristics and the degree of training of the raters. For example, it is probably more problematic to agree on "difficult" cases with unclear symptoms while extensively trained and calibrated raters could be expected to be more consistent in their ratings than uncalibrated clinicians using the instrument in everyday clinical practice. A high level of objectivity or interrater reliability is a prerequisite for high validity but does not automatically purport it. Thus, it is necessary to evaluate the objectivity of a standardized diagnostic instrument and to do it in same type of sample and setting that it will be used in.

The validity of a standardized diagnostic instrument refers to the extent that the results reflect for example correct diagnosis, presence of symptoms and symptom severity. There are also different forms of validity, for example face validity or whether the instrument appears to measure the intended construct and construct validity including convergent, divergent and discriminant validity that are linked to how a result correlates with other constructs. The most important validity aspect for diagnostic instruments though is diagnostic validity that designates how well the result of the instrument conforms to diagnosis. However, for ASD and psychiatry in general, a clinical diagnosis in itself is not something absolute and completely objective but the product of clinicians overall judgment concluding a diagnostic process. In the field, 'gold standard' clinical diagnosis is often used as a reference standard in studies of diagnostic validity drawing on the best estimate method (Gotham, Risi, Pickles, & Lord, 2007; S. H. Kim & Lord, 2012a; Mahoney et al., 1998), which involves a multidisciplinary team of expert clinicians that on the basis of the information from different sources and collected with different methods formulate a consensus diagnosis, called best estimate diagnosis or clinical consensus diagnosis. This could be seen as a variant of the Longitudinal, Expert, and All Data [LEAD] standard proposed by Spitzer in the 80ies (Spitzer, 1983). Using the result from an already existing measure is another example of a reference standard used to evaluate new diagnostic instruments. However, this procedure is methodologically problematic because it is impossible to claim that the result from an old instrument should be closer to the true diagnosis than that from the new instrument, and moreover that it seems inappropriate and even absurd to evaluate a new and supposedly better performing instrument with a less efficient instrument, something that also already Spitzer pointed out in his seminal paper presenting the LEAD standard (Spitzer, 1983). In summary, this has led to that in many studies applying the best estimate diagnosis as reference standard, the results of the studied instrument has also been used in the diagnostic decision-making to a various degree, i.e. the diagnosis is not completely independent from the instruments studied.

1.6.3 Standardized diagnostic interviews

In psychiatry in general, the diagnostic interview, aside from checklists, is the prototypic standardized instrument in research and clinical practice. The most common broadband interviews for children and adolescents, the Kiddie Schedule for Affective Disorders and Schizophrenia for School Aged Children - Present and Lifetime version [K-SADS-PL] (Kaufman et al., 2009; Kaufman et al., 1997) and the Development and Well-Being Assessment [DAWBA] (Goodman, Ford, Richards, Gatward, & Meltzer, 2000) include brief ASD sections, while the most widely used equivalents for adults, the Composite International Diagnostic Interview [CIDI] of the WHO (1994) and the Structured Clinical Interview for DSM-IV Axis I [SCID-I] (First, Spitzer, Gibbon, & Williams, 1997) lack such sections. More recent and popular examples of extensive diagnostic interviews for ASD are the Developmental, Dimensional and Diagnostic Interview [3di] (Skuse et al., 2004) and the Diagnostic Interview for Social and Communication Disorders [DISCO] (Nygren et al., 2009; Wing, Leekam, Libby, Gould, & Larcombe, 2002) operationalizing ICD-10/DSM-IV criteria for ASD while the shorter Asperger Syndrome (and high-functioning autism) Diagnostic Interview [ASDI] (C. Gillberg, Gillberg, Råstam, & Wentz, 2001) operationalizes the Gillberg and Gillberg criteria of Asperger disorder (I. C. Gillberg & Gillberg, 1989). However, the most psychometrically well-documented and widely used diagnostic interview in the field is probably the Autism Diagnostic Interview-Revised [ADI-R] (Rutter, Le Couteur, & Lord, 2003), often labeled 'gold standard' and included in many national clinical guidelines (NICE, 2011; SBU, 2013), and a common diagnostic tool in clinical practice of ASD around the world (Ashwood, Buitelaar, Murphy, Spooren, & Charman, 2015; de Bildt, Sytema, Zander, et al., 2015; Havdahl, von Tetzchner, Huerta, Lord, & Bishop, 2015; Lampi et al., 2010; Saemundsen, Magnusson, Smari, & Sigurdardottir, 2003; The National Board of Health and Welfare, 2009). The ADI-R is a comprehensive, investigator-based standardized caregiver interview designed for individuals in all ages to collect the necessary information to assess whether diagnostic criteria for autistic disorder and other ASDs according to ICD-10 and DSM-IV are met. The ADI-R is also largely consistent of DSM-5 although originally conceptualized according to ICD-10/DSM-IV criteria. The ADI-R has been and is used in a great variety of research and its (interrater) reliability and validity are well established in scientific settings by so-called research reliable interviewers, i.e. extensively trained and calibrated (see also Methods for details). It is also widely used but rarely studied in clinical practice.

1.6.4 Standardized diagnostic observation tools

Standardized observational instruments in psychiatry in general and for ASD in particular are much rarer than interviews. For ASD, the Psychoeducational Profile – Third Edition [PEP-3] (Schopler, Lansing, Reichler, & Marcus, 2005) that is basically used as a tool for developmental assessment and treatment planning designed for children between 2 and 7 years of age, also comprises a scale of which a part could be used to gather ASD diagnostically relevant information. However, the PEP-3 is rarely used as a diagnostic

instrument and published data on the diagnostic validity are very scarce. The most popular and widely used observational instrument, and perhaps of all diagnostic instruments for ASD assessment both in clinical practice and research, though, is probably the Autism Diagnostic Observation Schedule ([ADOS/-2], in the following the acronym ADOS will be used for both versions) (Lord, Rutter, DiLavore, & Risi, 1999; Lord et al., 2012). The use of the ADOS is recommended by the same guidelines as recommends the ADI-R (NICE, 2011; SBU, 2013; Volkmar, Siegel, et al., 2014), and is considered 'gold standard' in the field. The ADOS' owes its popularity for several reasons. Firstly, its "clinicalness" in that it samples and quantifies behaviors during a direct observation in an ongoing naturalistic interaction between the patient regardless of age and the observing examiner. Secondly, by its way of providing (diagnostic) ADOS classifications and severity scores to aid categorical and dimensional evaluation of ASD. Thirdly, its compatibility with the ADI-R. The ADOS can be administered to children from 12 months of age up to adults. The design of the ADOS largely follows the ICD-10/DSM-IV definition of the ASD and might be even more compatible with DSM-5 (Huerta, Bishop, Duncan, Hus, & Lord, 2012). The varying behavioral manifestations of ASD symptoms are controlled for by a language and age dependent module and algorithm system. There exists an extensive body of research establishing the interrater reliability (i.e. objectivity) and validity of the ADOS, mostly from research settings by research reliable examiners and regarding older children but fewer examples from clinical settings and for younger children.

1.6.5 Standardized instruments for assessment of impairment and functioning

It has been proposed that impairment in mental disorder should be assessed using standardized, normed and psychometrically sound instruments (Naglieri, 2009). As earlier mentioned, DSM-5 recommends the use of the standardized and normed WHODAS (Üstün et al., 2010) for this purpose. However, this instrument with excellent psychometric properties, translated into many languages and designed for use in different cultures, is still only available for adults, although there exists an unpublished version for school age children (Canino, 2013). The WHODAS draws on the ICF and assesses basically the ICF components activities and participation, which are closely related to the concept of adaptive functioning. Adaptive functioning is operationalized and measured by several standardized and normed instruments, for example the checklist Adaptive Behavior Assessment System, Second Edition [ABAS-2] (Harrison & Oakland, 2003), but it is the investigator-based caregiver interview Vineland Adaptive Behavior Scale [VABS/-II] (Sparrow, Balla, & Cichetti, 1984; Sparrow, Cichetti, & Balla, 2005) that probably is the most widely used instrument of this kind in the ASD field, both in research and clinical practice. Its reliability and validity as well as sensitivity for functioning difficulties in ASD are well established and it is normed from 0 to 90 years.

1.7 IMPLICATIONS OF THE USE OF STANDARDIZED DIAGNOSTIC INSTRUMENTS IN AUTISM SPECTRUM DISORDER

1.7.1 Assets

In research and clinical practice around the world, an increasing number of children, adolescents and adults with suspicion of ASD are assessed and/or diagnosed with the support of the standardized diagnostic instruments ADI-R and ADOS as well as getting their functioning assessed with the VABS. In general, the use of standardized diagnostic instruments with well-documented psychometric properties like the ADI-R and the ADOS holds the potential promise of improved consistency and reliability of symptom assessment across clinicians, time and geography (Spitzer, 1983; Wittchen, 1994). This use also lays the foundation of improved communication about ASD and comparability of results across settings and researchers. Large samples of consistently assessed and characterized individuals have been collected and standards for quantifying ASD symptomatology, studying dimensions, symptom severity and clustering of symptoms in a reliable way have been developed. For example, it has been convincingly demonstrated that the significant differences in the distribution of best estimate DSM-IV-TR based ASD subcategories (autistic disorder, Asperger's disorder and PDD-NOS) across sites had no foundation in quantified symptomatology differences, as measured by the ADI-R and the ADOS, but in different site specific diagnostic routines only (Lord et al., 2012). Furthermore, the influence of unspecific factors like age, IQ and language level on ASD symptomatology has also been studied in an unprecedented systematic and quantified fashion with the ADI-R (Hus & Lord, 2013; Hus, Pickles, Cook, Risi, & Lord, 2007) and the ADOS (de Bildt et al., 2009; Gotham et al., 2007; Oosterling et al., 2010). For the ADI-R, this has for example resulted in new research algorithms for toddlers with improved diagnostic validity (S. H. Kim & Lord, 2012b) and for the ADOS revised algorithms and a new module, the Toddler Module for children from 12 to 30 months without fluent speech (Gotham et al., 2007; Lord et al., 2012; Luyster et al., 2009). Moreover, based on large samples of well-characterized individuals with ASD, a severity metric of ASD symptomatology has been developed from ADOS domain and total scores with the result that the influence of language level and age has been reduced to a minimum: the Calibrated Severity Scores [CSS] (Esler et al., 2015; Gotham et al., 2009; Hus & Lord, 2014) also called Comparison Scores [CS] (Lord et al., 2012). Psychometrically sound comparisons of ASD symptom severity (as manifested during an ADOS administration) between individuals with different age and language level as well as intra-individual during development and module/algorithm change become possible as well as testing the hypothesis of the association between symptom severity and functional impairment (Kanne et al., 2011). This systematic and standardized way of assessing ASD symptomatology with the ADI-R and ADOS has also informed the new ASD definition of DSM-5, for example, the reorganization of the symptomatology in two clusters instead of three and the merging of the different DSM-IV-TR ASD diagnoses into only one diagnostic category. Moreover, this type of research has also contributed to the change of status of language delay from being a diagnostic criterion in DSM-IV-TR to become, together with IQ,

a specifier, i.e. important factors to take into account when assessing symptomatology but nonspecific to ASD (Lord, Corsello, & Grzadzinski, 2014). Even in regards to the new RDoC approach, where no such domain as ASD exists, the ADI-R's and the ADOS's unique and reliable way of assessing social and communicative behaviors might have the potential to become even a standard for sampling social communication in general.

1.7.2 Challenges

According to the DSM-5 and various guidelines, expert clinicians' judgment is still considered the diagnostic 'gold standard' for ASD and in psychiatry in general. However, the use of psychometrically sound standardized diagnostic instruments can improve the quality of such diagnostic decision-making (APA, 2013, p.55; SBU, 2013). Still, the availability of standardized diagnostic instrument like the ADI-R and the ADOS that generate numerical results and clear-cut diagnostic classifications, rather authoritative pieces of information, may also tempt the users to apply the results of these instruments straight off to supplant expert clinicians' judgments, something that has been discussed since the infancy of the history of standardized diagnostic instruments in psychiatry (Spitzer, 1983). It is not unusual in research to use the result of the ADI-R and the ADOS as a proxy for diagnosis (Bryson et al., 2007; Ozonoff et al., 2014; Wolff et al., 2012). Furthermore, in clinical practice, the result from the ADOS has begun to heavily influence diagnostic decision-making. This overreliance of the results of diagnostic instrument risks by misclassification to cause harm in clinical practice and bias in research. Ultimately, the risk of reification of diagnosis is imminent (Hyman, 2010). Clinicians and researcher must keep in mind that diagnostic instruments, operationalizing diagnostic criteria, must not be mistaken for the complete description of the disorder, as criteria are offered as guidelines only.

The authors of the ADI-R and the ADOS keep emphasizing that the proper use of these instruments require thorough clinical expertise of ASD besides the mastery of the instrument in question (Lord et al., 2012; Rutter et al., 2003), reminding that the published data on interrater reliability or objectivity and diagnostic validity (cutoffs, sensitivities and specificities) have been derived from research settings by research reliable experts. Objectivity and diagnostic validity of the ADI-R and the ADOS in clinical practice are largely unknown and despite their wide use in these settings we lack knowledge about how the estimates from research settings compare with those produced from clinical, non-research reliable users with variable experience.

It is also known that the diagnostic accuracy of the ADI-R and the ADOS, which is associated with the diagnostic definition, is lower for unclear cases, i.e. children younger than 48 months and older than 12 years, and/or high-functioning and fluently speaking individuals (Lord et al., 2014; NICE, 2011).

1.8   SUMMARY

Taken together, there are three important areas where we are in need for more knowledge regarding standardized diagnostic assessment of ASD: 1) objectivity or interrater reliability of the ADI-R and the ADOS in clinical practice; 2) diagnostic validity of the same instruments' revised algorithms in regards to toddlers and young preschoolers; 3) the effect on diagnosis rates of the newly introduced DSM-5 impairment criterion for ASD in regards to young preschoolers.

1.8.1   Objectivity of the ADI-R and the ADOS

Objectivity is a necessary but not sufficient prerequisite of diagnostic validity of an instrument. Excellent objectivity in the form of interrater reliability has been reported in several studies from research settings with highly trained, research reliable and consistently calibrated raters for both the ADI-R and the ADOS. To the author's best knowledge, no studies have reported data on objectivity in clinical practice settings with non-expert users with various levels of experience. This information is important to increase our understanding of the ADI-R and the ADOS psychometric properties in clinical practice and to what extent the database from research is applicable in clinical practice.

1.8.2   Diagnostic validity of the ADI-R and the ADOS in toddlers and young preschoolers

Diagnostic validity of the ADI-R and the ADOS in children below 4 years of age, an increasingly diagnosed and prioritized group, is less accurate and less studied in comparison to children 4 to 12 years of age with clear symptoms, some but not fluent language and some degree of intellectual disability (Lord et al., 2014; NICE, 2011). Previous research, especially in regards to the ADOS, has seldom focused on this specific age group but reported results from mixed age groups, not for all ASD versus non-ASD spectrum [NS] and have almost only been studied in US research settings. The standard algorithms of the ADI-R has shown to be either overinclusive (Lord, Storoschuk, Rutter, & Pickles, 1993) or underinclusive (Ventola et al., 2006; Wiggins & Robins, 2008) in young children. However, recently, Kim and Lord presented a new set of research algorithms for children between 12 and 47 months with improved diagnostic validity (S. H. Kim & Lord, 2012b; S. H. Kim, Thurm, Shumway, & Lord, 2013). Concerning the ADOS, previous research indicates that the ADOS, even in its revised form is overinclusive in regards to the youngest and/or most developmentally retarded children. Moreover, for the combined use of the ADI-R and the ADOS, applying the revised algorithms in children below 48 months of age (Gotham et al., 2007; S. H. Kim & Lord, 2012b; Lord et al., 2012), it exists only one study from a research setting in the US (S. H. Kim & Lord, 2012a).

1.8.3 The novel impairment DSM-5 criterion for Autism Spectrum Disorder in young children

In general, it has been recommended to use standardized and normed instruments to measure impairment (Bernier, 2012; Naglieri, 2009). However, DSM-5 does not provide an explicit definition of impairment even though it recommends the use of the standardized and normed WHODAS for adults, which is based on the ICF. As the ICF way of defining impairment and adaptive functioning are closely related, the use of the normed and standardized VABS appear to be a measure of choice of impairment in ASD, further motivated by its extensive use in the ASD field. To begin with, the VABS is a structured caregiver interview and normed for all ages. There exists an extensive knowledgebase of adaptive functioning concerning individuals with ASD showing substantially decreased levels for older individuals but a less clear picture for infants, toddlers and young preschoolers. Due to the properties of the construct, the naturally sheltered environments that young children live in, the low demand of adaptive skills and the higher thresholds to react on potential problems in the youngest children, it could be assumed that the new impairment criterion, depending on how it is defined, might affect diagnosis rates of ASD for young children, especially among those without intellectual and/or language impairment. Therefore, the introduction of the new impairment criterion in ASD might constitute a threat still not thought of to early diagnosis and early intervention by preventing young children with a manifest ASD symptomatology but not (yet) experienced as functionally impaired from fulfilling criteria for ASD. No study has examined the impairment criterion with standardized instruments or what the effects the application of different cutoffs for defining impairment might have on diagnosis rate for young children with ASD, a striking difference to the large amount of studies on the impact of the new symptomatology criteria on diagnosis rates. Few studies have examined the effect of symptom severity on impairment defined by the VABS, despite the assumption that symptoms cause impairment, or how the specifiers interact with impairment.

1.8.4 Conclusion

ASD is a public health concern demanding improved diagnostic methodologies, especially for younger children, and in clinical practice.

2 AIMS OF THE THESIS

The overall aim of this thesis is to study some previously under-researched psychometric properties of some of the internationally most widely used standardized diagnostic instruments for ASD, the ADI-R, the ADOS-2 and the VABS-II, focusing on the objectivity in clinical practice of the ADI-R and the ADOS operationalized as interrater reliability, the diagnostic validity for toddlers and young preschoolers of the same instruments as well as on the new DSM-5 impairment criterion in ASD in order to improve current diagnostic practice.

In studies I and II, the aim was to examine the objectivity of the ADI-R and the ADOS through the spontaneous interrater reliability on item level, for domain totals and classification as well as on criteria level for the ADI-R, across various naturalistic clinical settings among clinicians with different levels of clinical experience and expertise of using the instruments and from different clinical and research programs in Sweden.

In study III, the aim was to examine the diagnostic validity of the ADI-R and the ADOS separately and in combination in a Swedish clinical sample consisting of children below 48 months of age.

In study IV, the aim was to investigate the impact of the new DSM-5 impairment criterion on diagnosing ASD in toddlers and young preschoolers and how the new DSM-5 specifiers as well as age and gender were associated with impairment status

3 METHODS

3.1 PARTICIPANTS

3.1.1 Study I

Ten children and adolescents ($n = 8$ males) with suspected ASD were assessed by $n = 11$ raters to determine interrater reliability for the ADI-R. Nine participants with the mean age of 10 years (SD = 4.6 years, range = 2-17 years) had ASD ($n = 5$ Autistic Disorder, $n = 3$ Asperger's Disorder and $n = 1$ PDD-NOS, and one was typically developing (17 years old, verbal IQ [VIQ] = 100, nonverbal IQ [NVIQ] = 90). Mean VIQ and NVIQ were 104.6 (SD = 15.5) and 110.8 (SD = 7.3), respectively ($n = 5$).

3.1.2 Study II

Forty children and adolescents ($n = 28$ males) were examined with the ADOS modules 1-4, 10 of each module, and assessed by $n = 15$ raters for ADOS interrater reliability. In total, $n = 28$ had ASD ($n = 13$ autistic disorder, $n = 6$ Asperger's disorder and $n = 9$ PDD-NOS) and of the 12 participants without ASD, $n = 8$ had ADHD, $n = 1$ language disorder and intellectual disability respectively and $n = 2$ no diagnosis. Mean age for module 1 participants with ASD was 3.8 (SD = 1.0) and NS 3.7 (SD = 0.4), for module 2 with ASD 4.8 (SD = 0.7) and NS 4.2 (SD = 0.9), for module 3 with ASD 11.1 (SD = 2.1) and NS 9.4 (SD = 2.8), and for module 4 with ASD 16.0 (SD = 1.2) and NS 16.0 (SD = 2.5). Mean VIQ for participants with ASD and NS of module 1 was 60.3 (SD = 19.6) and 77.5 (SD = 6.4) and NVIQ was 88 and 77 (no SDs, only one participant) respectively; VIQ for module 2 was 89.3 (SD = 15.6) and 94.0 (no SD, only one participant) and NVIQ was not available; VIQ for module 3 was 92.3 (SD = 12.0) and 93 (no SD, only one participant) and NVIQ was 102.5 (SD = 16.4) and 98 (no SD, only one participant); and VIQ for module 4 was 97 (SD = 15.2) and 104.0 (SD = 12.7) and NVIQ was 94.2 (SD = 9.9) and 101.4 (SD = 14.2).

3.1.3 Study III

$N = 268$ children 18-47 months of age were assessed by a multidisciplinary diagnostic team with the ADI-R, the ADOS, IQ, adaptive functioning, etc. Mean age was 37.9 months (SD = 7.2 months). Following assessment, $n = 171$ children got an ASD diagnosis (autistic disorder: $n = 103$, PDD-NOS: $n = 68$) and $n = 97$ children were classified as NS: $n = 67$ children received a non-ASD neurodevelopmental diagnosis (intellectual disability: $n = 9$, attention deficit hyperactivity disorder (ADHD): $n = 16$, language disorder: 42), and $n = 30$ children no psychiatric diagnosis but in most cases showing different kinds of special needs due to developmental delays and adaptive and behavioral problems that were too subtle or vague for qualify for a diagnosis. The majority of the participants were boys (76%) and of Swedish ethnicity (59%), i.e. born by a mother of that ethnicity. Neither gender nor ethnicity was associated significantly with diagnosis. Sample characteristics by ADI-R and ADOS algorithms are presented in Tables 1 and 2.

Table 1. Sample description by ADI-R developmental cells

	12-20/NV21-47				SW21-47				PH21-47			
	ASD	NS	F	p	ASD	NS	F	p	ASD	NS	F	P
Gender (male, female)	41, 18	11, 2			46, 10	20, 12			39, 8	34, 13		
Age in months (SD)	32.4 (7.0)	33.4 (9.6)	.20	.655	38.6 (4.8)	38.9 (7.3)	.04	.837	41.9 (4.2)	41.9 (5.2)	.00	.983
Mean NVIQ (SD)	67.0 (19.0)	79.2 (24.3)	3.57	.064	73.3 (14.0)	84.8 (17.0)	10.98	.001	86.8 (10.9)	94.3 (13.4)	7.53	.007
Mean VABC (SD)	67.8 (7.8)	81.1 (7.6)	31.08	<.001	74.8 (8.7)	82.09 (8.2)	14.66	<.001	77.2 (8.1)	84.2 (9.9)	13.95	<.001
ADI-R SA/SC	9.7 (4.2)	4.6 (4.0)	15.46	<.001	8.6 (4.3)	5.1 (3.8)	15.00	<.001	7.9 (4.7)	4.3 (3.7)	16.74	<.001
ADI-R RRB	3.0 (2.2)	1.8 (1.7)	3.80	.055	3.0 (2.5)	1.5 (1.4)	9.26	.003	2.8 (2.5)	1.9 (2.4)	3.59	.061
ADI-R IPG/RPI	8.8 (2.8)	4.5 (3.2)	24.70	<.001	5.3 (2.7)	3.4 (2.2)	15.72	<.001	3.5 (2.0)	2.0 (1.8)	14.57	<.001
ADI-R total	12.7 (5.6)	6.4 (5.1)	13.83	<.001	11.6 (6.1)	6.6 (4.2)	16.69	<.001	14.2 (7.4)	8.2 (6.2)	18.27	<.001
ADOS SA	16.9 (3.0)	8.8 (4.9)	54.70	<.001	13.0 (3.9)	6.1 (4.6)	54.87	<.001	12.8 (3.2)	5.5 (3.3)	120.30	<.001
ADOS RRB	3.7 (2.0)	1.4 (1.4)	14.52	<.001	2.5 (1.8)	1.0 (1.1)	18.50	<.001	2.4 (1.8)	0.7 (0.8)	33.39	<.001

NVIQ = Non Verbal IQ, VABC = Vineland-II Adaptive Behavior Composite, ADI-R SA/SC = Autism Diagnostic Interview-Revised Social Affect/Social Communication, ADI-R RRB = Repetitive and Restricted Behavior, ADI-R IGP/RPI = Imitation, Gestures, and Play/Reciprocal Peer Interaction, ADOS SA = Autism Diagnostic Observation Schedule Social Affect, ADOS RRB = Repetitive and Restricted Behavior, 12–20/NV21– 47 = all children 12–20 months + nonverbal children 21–47 months, SW21–47 = children 21–47 months with single words, PH21–47 = children 21–47 months with phrase speech. In the 12-20/NV21-47 group, three children were less 21 months old and NVIQ was available for 43/59 children; SW21-47: NVIQ was available for 49/56 children; PH21-47: NVIQ available for 37/47 children with ASD and VABC for 46/47 children.

Table 2. Sample description by ADOS module and algorithm.

	ADOS Module 1 few to no words						ADOS Module 1 some words						ADOS Module 2, younger than 5 years					
	N	Mean	SD	Range	F	p	N	Mean	SD	Range	F	p	N	Mean	SD	Range	F	p
ASD																		
Gender (m, f)	58, 14						50, 14						25, 7					
Age	72	33.6	7.2	21-47	1.3	.25	64	38.3	5.7	20-47	.3	.566	32	42.6	3.6	34-47	.3	.603
VIQ	39	42.6	27.6	6-94	20.1	<.001	41	60.4	21.6	10-94	8.5	.005	29	85.6	10.0	65-100	1.7	.199
NVIQ	52	68.1	17.6	23-94	9.1	<.001	52	76.2	14.5	45-109	6.0	.016	28	88.2	10.0	65-112	6.2	.015
VABC	72	68.7	7.3	51-83	42.6	<.001	63	74.8	9.4	53-100	5.0	.028	32	78.6	7.4	66-95	13.9	<.001
ADOS SA	72	16.6	2.8	9-20	66.7	<.001	64	12.7	3.8	5-20	67.7	<.001	32	12.5	3.2	6-19	87.7	<.001
ADOS RRB	72	3.4	2.1	0-8	13.7	<.001	64	2.4	1.5	0-5	29.7	<.001	32	2.5	1.9	0-8	26.0	<.001
ADOS SA + RRB	72	20.1	3.9	10-27	76.5	<.001	64	15.1	4.3	6-24	81.9	<.001	32	15.0	3.8	7-22	109.0	<.001
NS																		
Gender	18, 3						15, 14						29, 11					
Age	21	35.7	7.9	20-45			29	39.0	6.5	25-47			40	43.1	4.4	29-47		
VIQ	20	72.4	15.1	46-102			28	73.9	14.0	40-100			40	90.5	17.6	44-139		
NVIQ	21	81.1	14.2	65-110			29	84.5	15.0	39-114			40	95.5	12.9	71-132		
VABC	21	80.7	7.8	65-97			29	79.3	8.2	66-102			40	86.0	9.2	65-102		
ADOS SA	21	9.8	4.9	4-17			29	5.7	3.8	1-15			40	5.2	3.4	0-14		
ADOS RRB	21	1.5	1.3	0-4			29	0.8	1.0	0-4			40	0.8	0.8	0-3		
ADOS SA + RRB	21	10.9	5.2	4-20			29	6.4	4.2	1-19			40	5.9	3.8	0-15		

Gender (m, f) = male, female; VIQ = Verbal IQ; NVIQ = Non Verbal IQ; VABC = Vineland-II Adaptive Behavior Composite; ADOS SA = Social Affect; ADOS RRB = Repetitive and Restricted Behavior; ADOS SA + RRB = Total score; Nine children in module 1 NW had a NVMA of 15 months or below.

3.1.4 Study IV

$N = 171$ children aged 20–47 months with ASD, a subsample of the sample of study III, were assessed comprehensively by a multidisciplinary diagnostic team. Following assessment, $n = 127$ (77 % boys) children ($n = 68$ with autistic disorder and $n = 59$ with PDD-NOS) were included in the analyses. Inclusion criteria were, besides a clinical DSM-IV-TR diagnosis of ASD, available assessment results from cognitive testing for NVIQ, the ADI-R, the ADOS and the VABS-II Survey Interview. See Table 3 for more details.

Table 3. Sample characteristics.

	Mean (SD)	Range
Age at diagnosis (months)	37.7 (6.7)	20–47
NVIQ	75.4 (16.7)	23–112
VIQ	61.3 (27.3)	6–100
Vineland-II ABC	73.7 (8.7)	54–100
Communication	73.8 (13.7)	42–113
Daily Living Skills	77.8 (9.8)	52–107
Socialization	72.1 (9.0)	53–105
Motor Skills	83.6 (10.6)	56–111
ADOS-2 SA CS	7.3 (1.8)	3–10
ADOS-2 RRB CS	6.4 (2.3)	1–10
ADOS-2 Total CS	7.0 (1.7)	2–10
ADI-R		
12-20/21-47 months Non verbal ($n = 41$)	12.2 (5.2)	2–22
21-47 months Some words ($n = 49$)	11.9 (6.1)	1–27
21-47 months Phrases ($n = 37$)	12.8 (6.9)	2–26

Note. N = 127, (n= 29 girls, n = 98 boys); NVIQ = Non Verbal IQ; VIQ = Verbal IQ; Vineland-II ABC = Vineland Adaptive Behavior Scales, 2nd edition, composite score; ADOS-2 SA CS = Autism Diagnostic Observation Schedule, 2nd edition, Social Affect Comparison Score; ADOS-2 RRB CS = Restricted and Repetitive Behavior; ADI-R = Autism Diagnostic Interview-Revised; 12-20/21-47 months Non-verbal = research algorithm of the ADI-R for children aged 12-20 and non-verbal children aged 21-47 months; 21-47 months Some words = children aged 21-47 months and using some words functionally; 21-47 Phrases = children aged 21-47 months using functional phrase speech.

3.2 PROCEDURE

3.2.1 Study I

Participants were recruited from clinical routine of two child and adolescent psychiatry outpatient departments in Sweden between 2011 and 2014. Their caregivers were administered the ADI-R, and video-recorded, by a research reliable interviewer employed at the units as part of a regular clinical routine diagnostic evaluation. Ten psychologists and one pediatrician from 7 different clinical outpatient units and one clinical research center from all over Sweden participated in the study. Owing to the influx rhythm of video recorded interviews and the availability of raters, the eleven raters scored between 1 and 9 interviews. ADI-R expertise and experience of using the interview varied substantially. Four of the 11 raters were research reliable and certified ADI-R trainers, the other seven raters had basic clinical training on the instrument. Each of the ADI-R video recorded administrations on the 10 participants was rated by five independent raters, i.e. 50 rated ADI-Rs in total. Raters were blind for diagnostic status in all but 4 ratings, which were the scores from the original ADI-R clinical assessment administration. One brief initial calibration meeting was the only preparation prior to the current study of ADI-R objectivity, as the intention of this research was to investigate the spontaneous interrater reliability of the ADI-R in everyday clinical use.

3.2.2 Study II

Participants were recruited from clinical routine of several child and adolescent psychiatry outpatient departments or specialized neuropsychiatric units in Sweden between 2011 and 2014 and administered a videotaped ADOS as part of regular clinical routine diagnostic evaluation. Fourteen psychologists and one pediatrician from 13 different clinical centers and one research center from all over Sweden participated in the study. The combination of inflow of video recorded administrations and disposable raters, the 13 clinicians rated between 1 and 8 administrations of modules 1 and 2, respectively, and between 1 and 7 administrations in module 3, while 15 clinicians rated between 1 and 6 administrations of module 4. ADOS expertise and experience of administering the measure varied substantially. Three of the 15 raters were research reliable and certified ADOS trainers, the other 12 raters had attended ADOS basic clinical training. Each of the in total 40 videotaped ADOS administrations was rated by five independent raters, i.e. 50 ratings for each module. Four of five raters were blind for diagnostic status of the individual examined with the ADOS, while the fifth rater was the clinician who originally had examined the participant in clinical practice. Two brief calibration meetings prior to the study were scheduled, as the intention was to investigate the spontaneous or "true" interrater reliability of the ADOS in everyday clinical use.

3.2.3 Studies III and IV

Participants were referred to the Neuropsychiatric Resource Team Southeast, Child and Adolescent Psychiatry, Stockholm County Council between 2006 and 2012 due to unclear

developmental concerns, for instance, language delay or global developmental delay, interaction difficulties, internalizing or externalizing behavior problems. The unit is a multidisciplinary diagnostic specialist outpatient clinic and part of the public health care serving preschool children below 48 months of age (~ 60 children/year). All children underwent a comprehensive developmental assessment routine by a multidisciplinary team consisting of child psychiatrists, psychologists, and social workers including history taking by a child psychiatrist, a psychologist testing the child's cognitive abilities and assessing his/her adaptive functioning through the Vineland-II survey interview. Furthermore, one of the team members spent half a day observing the child at his/her preschool as well as interviewing the staff. The ADOS was administered by a formally clinically trained or research reliable examiner and routinely observed by an additional ADOS trained team member, both coding independently and reaching consensus after discussion. The ADI-R was administered and coded by any of the team members not being familiar with the child and not taking part in the diagnostic clinical consensus discussion. Two of the authors of this study are certified ADI-R and ADOS trainers. All available information, results and observational data from the assessments were discussed by the team having seen the child and his or her parents to generate a clinical consensus diagnosis according to DSM-IV-TR.

In study IV, the ADOS and the ADI-R were used to assess the presence of the ASD symptomatology according to DSM-5 following the same method as Huerta et al. (2012). Relevant items from both instruments were combined and plotted on the ASD criteria and the presence of a symptom was defined as a score of at least one on a relevant item on any of the instruments. For criterion A, indices for the three groups of the criterion were created by combining items from the ADOS and the ADI-R (A1-3: deficits in social-emotional reciprocity, deficits in nonverbal communicative behaviors and deficits in developing, maintaining, and understanding relationships). To fulfill the A criterion of DSM-5, documented symptoms (score $\geq$ 1) on any of the three groups were required as it is not clear that all three parts of the A criterion are required to fulfill the criterion (''by the following'') according to DSM-5. For the B criterion (restricted and repetitive behavior), the same type of indices as for the A criterion were created for the four different groups of the criterion (B1-4: stereotyped and repetitive behaviors, insistence on sameness, strong interests and sensory issues). To fulfill the B criterion of DSM-5, documented symptoms from at least two of the four groups were required. Ninety-one percent of the children (115/127) exhibited symptoms on the ADOS and/or the ADI-R to fulfill the DSM-5 A and B criteria. The three parts of criterion A (clear symptoms documented by the ADOS and/or the ADI-R) were fulfilled by all 127 children. Criterion B (clear symptoms documented by the ADOS and the ADI-R from at least two of the four parts of the criterion) was fulfilled by 115 children; $n = 10$ had one symptom ($n = 8$ with PDD-NOS) and $n = 2$ had none (both had PDD-NOS).

3.3 MEASURES

3.3.1 The ADI-R and language

In studies I, III and IV, the Swedish translation of the ADI-R (Rutter et al., 2003) was administered. The ADI-R is a standardized, investigator-based diagnostic caregiver interview developed in research settings in the US, Canada and the UK and translated into almost 20 languages. The majority of the ADI-R items cover specific behaviors associated with ASD and operationalize the diagnostic criteria of infantile autism/autistic disorder according to ICD-10/DSM-IV. The standard version was designed and validated for individuals with a mental age of at least 24 months up to adulthood to be used in the evaluation of suspected ASD. There are now even revised diagnostic algorithms of the ADI-R available for children from 12 to 47 months of age with a mental age of at least 10 months (S. H. Kim & Lord, 2012b). The ADI-R comprises 93 items and 153 ratings (most items are scored for "current" as well as "ever" or "most abnormal 4-5 years" period) and is organized in six sections: early development, acquisition and loss of language/other skills, language and communication functioning, social development and play, interests and behaviors and general behaviors. Each of the 12 behavioral diagnostic criteria of ASD in ICD-10/DSM-IV are operationalized in the ADI-R by 2 to 5 items resulting in a relatively time consuming measure, administration time is typically 90-180 minutes, but also potentially reliable (Rutter et al. 2003, p. 40). The interviewer elicits detailed behavioral descriptions of the individual in question from the caregiver through a set of verbatim questions and prompts to use. The items are typically scored from 0 to 3, where "0" indicates that a specified behavior is/was not present, scores of "2" and "3" reflect definite prototypic and severe ASD symptomatology, and "1" prototypic but mild symptomatology. The same interview questions are utilized regardless of age, but to control for age and language effects described earlier on ASD symptomatology, the assessment of symptoms are made using information for different time periods (most often 4-5 years) and different diagnostic algorithms are applied according to the individual's age and language level. The standard diagnostic algorithms of the ADI-R are made up of combinations of 33 to 42 of the most diagnostically discriminating items/ratings of the interview depending on age and language level of the patient. They are organized in four sections consistent with the diagnostic definition of autism in ICD-10/DSM-IV and have separate diagnostic cutoffs for each of the sections: Qualitative abnormalities in reciprocal social interaction (RSI), qualitative abnormalities in communication (COM), restricted and repetitive patterns of behavior (RRB) and abnormality of development evident at or before 36 months. The three new research algorithms for toddlers combine 13 to 20 items, also according to age and language level, and are composed of three sections: Social Affect (SA)/Social Communication (SC), Restricted and Repetitive Behavior (RRB), and Imitation, Gestures, and Play (IGP)/Reciprocal Peer Interaction (RPI). They have one single diagnostic cutoff for the algorithm total, which makes the structure of these algorithms consistent with the revised ADOS-2 algorithms and with DSM-5.

In study III, the new research algorithms (S. H. Kim & Lord, 2012b) were applied. The

different research algorithms depend on the child's age and expressive language: children aged 12–20 months and those aged 21–47 months without speech (12-20/21-47NV), children aged 21–47 months using single words (SW21-47), and children aged 21–47 months with developed phrase speech (PH21-47). For each algorithm, different item combinations are used to generate a single total. For the NV and SW groups, only the SA and RRB domains are combined to generate totals (NV: 13 items and SW: 16 items), while for the PH group, the SC, RRB and RPI domains (20 items) are combined to form the algorithm total. There is a clinical lower cutoff, optimizing sensitivity, and a higher research cutoff, optimizing specificity, as well as ranges of concern (little-to-no, mild-to-to moderate, and moderate-to-severe). In study III, $n = 72$ participants fell into the NV group (three of them were aged < 21 months), $n = 88$ into the SW group, and $n = 94$ into the PH group. In the NV group, VIQ correlated significantly with the different ADI-R domain totals ($r = -.45$ to $-.50, p < .005$) and age correlated with both the total score and the SA domain score ($r = -.34, p \leq .004$). No other participant characteristics (age, VIQ, and NVIQ) and domain totals correlated significantly or $r > .40$.

In study IV, the assessment of the DSM-5 specifier language impairment was made based on item 30 of the ADI-R in which the child's current level of functional language is evaluated. Children between 20 and 35 months of age had to use more than five different words functionally on a daily basis (item scores = 0 and 1) while children between 36 and 47 months of age had to use functional phrase speech (item score = 0) in order to be considered having typical language development. According to this evaluation $n = 76$ children (60 %, $n = 56$ boys) were estimated having language impairment.

The psychometric properties of the ADI-R are well-documented in research contexts among highly trained so-called research reliable raters. The diagnostic validity of the ADI-R have been evaluated in different age groups in many studies, with reported sensitivities and specificities between 80 and to over 90% (de Bildt et al., 2013; de Bildt et al., 2004; Gray, Tonge, & Sweeney, 2008; Le Couteur, Haden, Hammal, & McConachie, 2008; Le Couteur et al., 1989; Lord et al., 1997; Lord, Rutter, & Le Couteur, 1994; Mazefsky & Oswald, 2006; Papanikolaou et al., 2009; Sappok et al., 2013). For toddlers, using the clinical cutoffs of the new research algorithms, Kim and colleagues reported sensitivities for ASD versus NS disorders between 85% and 90% in the NV group, 94%–97% in the SW group, and 80%–89% in the PH group, while specificities ranged from 64% to 94% in the NV group, from 58% to 83% in the SW group, and from 70% to 94% in the PH group (S. H. Kim & Lord, 2012b; S. H. Kim et al., 2013), while de Bildt et al. (2015) in a large European multicountry study ($N = 1104$) reported sensitivities from 72% to 82% in the NV group, 73%–85% in the SW group, and 48%–64% in the PH group, and specificities from 84% to 98% in the NV group, from 59% to 79% in the SW group, and from 73% to 87% in the PH group. Studies on high-functioning adolescents (Gilchrist et al., 2001) and adults with ASD (Howlin, Moss, Savage, & Rutter, 2013) have also demonstrated that the ADI-R accurately identifies the majority of individuals with ASD, but without providing sensitivities and specificities. The factorial validity of the ADI-R has been explored and yielded different three factor solutions

explaining 38-50.5% of the item variance (Bölte & Poustka, 2001; Lecavalier et al., 2006; Tanguay, Robertson, & Derrick, 1998). Internal consistency of ADI-R domains ranged from a Cronbach's α of .79 to .95 for RSI, .76 to .84 for COM, and from .35 to .69 for RRB (Lecavalier et al., 2006; Lord et al., 1994). Finally, test-retest reliabilities of algorithm items have also been reported to be κ_w [weighted kappa] = .72 on average by Lord and colleagues (Lord et al., 1994), and κ_w = .49 for RSI, .51 for COM and .40 for RRB items in Hill et al. (Hill et al., 2001), and in the latter study, stability ranged from κ = .82 to .91 for domain cutoffs and was κ = .76 for overall diagnostic classification.

In regards to objectivity or interrater reliability, excellent results on item level has been reported for the ADI-R in several studies with κ_w often above .60 and percent exact agreement [PA] exceeding 90% (Cicchetti, Lord, Koenig, Klin, & Volkmar, 2008, 2014; Hill et al., 2001; Howlin et al., 2013; Le Couteur et al., 1989; Lord et al., 1994; Lord et al., 1993; Poustka et al., 1996; Tsuchiya et al., 2013). For diagnostic criteria as they are operationalized by the ADI-R, the interrater reliability is generally excellent with Intraclass Correlation [ICC] coefficients of .93 to .96 (Lord et al., 1994), .82 to .96 (Tsuchiya et al., 2013) and .52 to .89 (Poustka et al., 1996), as well as for domain totals where ICCs range from .94 to .97 (Le Couteur et al., 1989) and .59 to .87 (Chakrabarti & Fombonne, 2001). The interrater reliability of diagnostic classification has been reported by Chakrabarti and Fombonne (Chakrabarti & Fombonne, 2001): 81.6-92.1% and Hill et al. (Hill et al., 2001): κ = .74-.82.

3.3.2 The ADOS and symptom severity

The Swedish translation of the ADOS (Lord et al., 1999) applying the revised algorithms included in the ADOS-2 (Gotham et al., 2007; Lord et al., 2012) was administered in studies II-IV. The ADOS is a standardized expert direct observation scale developed in the US, Canada and the UK but used internationally and translated into at least 20 languages. It is designed to sample important social-communicative behaviors as well as any stereotypic and repetitive behavioral features according to the ICD-10 and DSM-IV/-TR/-5 criteria for autism/ASD in an ongoing semi-structured interaction between the assessed person and the examiner. The ADOS is appropriate for individuals with a suspected ASD from a mental and chronological age of at least 12 months up to adults through its systems of 5 different modules (sets of activities and tasks) and 8 algorithms (scoring procedures) designed to minimize the influence of language and age on the manifestation of ASD symptomatology. The administration of the ADOS typically takes around 45 minutes. Module 1 is conceived for children without fluent phrase speech, and two different algorithms are used depending on the expressive language level: one if the child uses few to no words (module 1 few to no words), the other if the child uses some words (module 1 some words). For children between 12 and 30 months of age on the same expressive language level, the new toddler module is now available (Lord et al., 2012; Luyster et al., 2009). Module 2 fits for children who use phrases consistently but not yet in a complex way (below the expressive language level of a typical 4 year old). Module 2 has two algorithms as well but this time depending on the child's age: below 5 years or 5 years and older. Module 3 and 4 is designed for children,

adolescents and adults with fluent and complex speech. Module 3 is recommended from 3 to approximately 16 years and module 4 from younger adolescence to adulthood. Both modules are scored using one algorithm each. Each criterion of the ICD-10/DSM-IV, according to the operationalization of Huerta and colleagues (2012) is represented by 1 to 7 items on the ADOS but four of the 12 criteria lack representation at all on the ADOS, mostly for modules 1 and 2, while for DSM-5 only two criteria for modules 1 and 2 lack. All items of the ADOS are typically coded from 0 (no abnormality related to autism/as specified) to 2 (definite evidence of abnormality) and sometimes 3 (profound severity). In the revised ADOS/ADOS-2 algorithms, items from the Social Interaction and Communication domains have been restructured to form the SA domain. The RRB domain, not part of the original algorithm scoring, has been added. Each module consists of 29 to 41 items from which a selection of the 14 most diagnostically informative items form the algorithm except for the revised module 4 that has 15 items (Hus & Lord, 2014). The algorithm totals are compared against cutoffs for the ADOS/-2 classifications of autism and autism spectrum except for the toddler module that has three different "ranges of concern" instead of an ADOS-2 classification. An ADOS/-2 classification is not necessarily equivalent to a clinical diagnosis.

In study II, the totals and the classifications for all algorithms of modules 1-3 of the ADOS-2 were used in the analyses as well as those of the revised algorithm for module 4 (Hus & Lord, 2014) but not the toddler module. Only the lower autism spectrum classification cutoff of the ADOS was used in the analyses, i.e. all individuals with any ASD as a group were tested against the autism spectrum cutoff.

In studies III and IV, the toddler module, module 1 (few to no words and some words algorithms), module 2 (younger than 5 years algorithm), and module 3 (study III) were applied. In study III, $n = 5$ children were given the toddler module, $n = 93$ module 1 few to no words algorithm, $n = 93$ module 1 some words algorithm, $n = 72$ module 2 (< 5 years algorithm), and $n = 2$ module 3. The correlations between age, VIQ and NVIQ and ADOS domain totals were $r \leq -.40$ in most cases. In module 1 few to no words, the correlation of VIQ and NVIQ and the ADOS total score was $r = -.47$ ($p < .001$) and in module 1 some words, VIQ and the ADOS total and SA domain totals was $r = -.41$ ($p < .001$) in both cases. The correlation between age and domain totals was $r \leq -.21$ in all modules.

In study IV, $n = 48$ children were administered a module 1 few to no words algorithm, $n = 51$ module 1 some words algorithm, and $n = 28$ module 2 below 5 years algorithm. Symptom severity was measured with the ADOS-2 CS (Hus, Gotham, & Lord, 2014; Lord et al., 2012). The CS is based on the totals of the SA and RRB domains and quantifies ASD symptom severity as it appears during the ADOS-2 administration corrected for age and language level. CS ranges from 1 (low) to 10 (high), with scores of 1–3 being below the autism spectrum cutoff, and scores of 4–10 being above the autism spectrum cutoff.

The psychometric properties of the ADOS are well-documented and a variety of studies, most often from research contexts among highly trained research reliable examiners, have reported on diagnostic validity in different mixed groups with sensitivities and specificities around

26

90% and above (Bastiaansen et al., 2011; Bölte & Poustka, 2004; de Bildt, Sytema, Meffert, & Bastiaansen, 2015; de Bildt et al., 2009; Gotham et al., 2008; Gotham et al., 2007; Gray et al., 2008; Hus & Lord, 2014; S. H. Kim & Lord, 2012a; Klein-Tasman, Risi, & Lord, 2007; Le Couteur et al., 2008; Lord et al., 2000; Oosterling et al., 2010). The revised algorithms of module 1 and 2 included in ADOS-2 and used in study III and IV have been evaluated in two large US research samples (Gotham et al., 2008; Gotham et al., 2007) and two Dutch research samples (de Bildt et al., 2009; Oosterling et al., 2010). In Gotham et al. (2008; 2007), sensitivities for module 1 (few to no words and some words) and module 2 (< 5 years of age algorithm) ranged from 86% to 98%, whereas specificities were 80% to 100% using the autism cutoff for autism only versus NS. The diagnostic validity for non-autism ASD versus NS was generally lower. In the Dutch samples, specificities and particularly sensitivities were generally lower than in the US studies, and Oosterling et al. (2010) reported for example sensitivities of 61-93% and specificities of 70-86% for ASD versus NS for module 1 (few to no words and some words) and module 2. The revised algorithms have also been used in clinical settings, for instance by Gray et al. (2008) in Australia for the comparison between autism versus non-autism ASD plus NS using the autism cutoff where sensitivities for module 1 (few to no words and some words) ranged between 89% and 98% and specificities between 82% and 86%, thus comparable to those found in US research samples for the comparison autism excluding non-autism ASD versus NS. ASD versus NS using the autism spectrum cutoff yielded somewhat lower sensitivities (78-92%) and specificities (86-92%). In a US clinical sample using the autism cutoff for the comparison autism versus NS, Molloy et al. (2011) reported sensitivities of 63-83% and specificities 65-81%, while sensitivities for ASD versus NS were 76-98% and specificities 29-60% for module 1 few to no words and some words as well as module 2 younger than 5 years.

All studies from both research settings and clinical practice that evaluated the performance of modules 1-2 examined children of a large age range (13-144 months) without an explicit separate analysis of young children, which makes the diagnostic accuracy of the ADOS in children aged less than 48 months basically unknown. The validity of the ADOS has also been examined with exploratory and confirmatory factorial analyses, which have yielded a two factor solution (SA and RRB) as the most efficient way of explaining the item variance (Gotham et al., 2007, 2008; Hus & Lord, 2014; Oosterling et al., 2010). Internal consistency of ADOS domains ranged from a Cronbach's α of .84 to .92 for SA and .51 to .66 for RRB for modules 1-3 (Gotham et al., 2007; Hus & Lord, 2014) and for module 4, .75 and .85 for communication and reciprocal social interaction respectively as well as .47 for RRB (Lord et al., 2012). Test-retest reliabilities for domain totals on average for modules 1-3 have been reported to be ICC = .89 for SA, .74 for RRB and .90 for overall totals (Lord et al., 2012).

In regards to objectivity or interrater reliability, Lord and colleagues have demonstrated the feasibility of reaching substantial levels of interrater reliability for groups of well-prepared raters on item level already in regards to the pre-published versions of the ADOS (κ_w between .58 and .92) (Lord et al., 1989) and the PL-ADOS (κ_w = .60-1.00) (DiLavore, Lord, & Rutter, 1995). The most comprehensive interrater reliability study of the ADOS to date was

conducted during the development of the first published version of the ADOS (Lord et al., 1999). Twelve raters were assessing $n = 98$ individuals ($n = 20\text{-}29$ per module) on item level and the interrater reliability was analyzed using κ_w and PA. For module 1, all but one item had a $\kappa_w > .60$ ($\kappa_w = .55\text{-}1.00$, median $= .78$), most items of module 2 had a $\kappa_w > .50$ ($\kappa_w = .38\text{-}.93$, median $= .65$), and many items of module 3 and 4 had a $\kappa_w > .60$ (module 3: $\kappa_w = .46\text{-}1.00$, median $= 61$; module 4: $\kappa_w = .41\text{-}.93$, median $= 60$). The interrater reliability of domain totals for Social Interaction was ICC $= .93$, .84 for Communication, .92 for Reciprocal Social Interaction and Communication combined and .82 for Stereotyped Behaviors and Restricted Interests for all modules pooled. Classification was assessed using PA: Module 1: 93%, module 2: 87%, module 3: 81% and module 4: 84% exact agreement when all participants (both with autistic disorder and PDD-NOS) were included, but PA was higher if individuals with PDD-NOS were excluded (90-100%). Additional interrater reliability data from another subsample of the dataset of 1999/2000 and the same 12 raters for domain totals and classification of the revised algorithms have been published in the ADOS-2 manual (Lord et al., 2012). ICC for SA of module 1 was .97 ($n = 63$), .98 for module 2 ($n = 50$) and .92 ($n = 66$) for module 3. RRB had ICCs of .79, .80 and .91 for modules 1, 2 and 3, while the ICCs of the overall totals were .97, .96 and .94. The interrater reliability of classification for modules 1 to 3 from still another subsample of the same dataset as above was 95% for module 1 ($n = 46$ autism; $n = 13$ non-spectrum), 98% for module 2 ($n = 28$ autism; $n = 6$ non-spectrum) and 92% ($n = 46$ autism; $n = 1$ autism spectrum). All the twelve contributing examiners in this study were thoroughly trained, had reached research reliability and also took part in weekly coding meetings with continuous and systematic checks of the interrater reliability. Apart from the ADOS authors, Bölte and Poustka (2004) have reported on interrater reliability from a genetics research project in Germany. Twelve individuals with autistic disorder (three for each module) were independently assessed by five raters and the interrater reliability of classification was $\kappa_w = 1.00$. In the Netherlands, Bastiaansen and colleagues (2011) evaluated the interrater reliability of module 4 in a sample of $n = 38$ with high-functioning ASD, $n = 18$ with schizophrenia, $n = 16$ with psychopathy, and $n = 21$ with typical development ($N = 93$). Each participant was rated by two raters from a pool of five research reliable psychologists, continuously calibrating themselves (de Bildt, Sytema, Meffert, et al., 2015). Mean κ_w was .66, with $\kappa_w > .60$ for 14 of the 21 items and none $\kappa_w < .50$. The interrater reliability for domain totals ranged from ICC $= .79$ (Communication) to .92 (Reciprocal Social Interaction as well as the overall total). The interrater agreement of classification was $\kappa = .73$ and PA $= 89.2\%$ using the lower autism spectrum cutoff and dividing the sample dichotomously in ASD/non-ASD groups.

3.3.3 IQ and intellectual disability

In all studies, Merrill-Palmer-R (Roid & Sampers, 2005), Wechsler Preschool and Primary Scales of Intelligence–Third Edition (Wechsler, 2004), or Mullen Scales of Early Learning (Mullen, 1995) were used to measure NVIQ and VIQ. In study I and II, Wechsler Intelligence Scale for Children-Fourth Edition (Wechsler, 2003) and Wechsler Adult Intelligence Scale-Fourth Edition (Wechsler, 2008) was also used. In study III and IV, VIQ and NVIQ were

calculated by averaging the age equivalents of the nonverbal subtests (Visual Reception and Fine Motor (Mullen) and Cognitive and Fine Motor (Merrill-Palmer-R)), and the verbal subtests to obtain mental age, which was divided by chronological age and multiplied by 100. In study IV, these tests and procedure were used to measure intellectual impairment that was defined as NVIQ < 70. Previous studies showed that IQ and adaptive functioning in individuals with ASD were strongly positively associated (Bölte & Poustka, 2002; Klin et al., 2007). In the current study, $n = 44$ (35 %) of the children ($n = 33$ boys) scored in the intellectual impairment range, which is proportional to the distribution of IQ in ASD in the region of the study (Idring et al., 2015).

3.3.4 Adaptive functioning and impairment

To measure adaptive functioning in studies III and IV, the VABS-II Survey Interview (Sparrow et al., 2005) was used, a structured caregiver interview that assesses age appropriate self-sufficiency skills. It provides norms across a large age range (birth to 90 years), has demonstrated excellent reliability and validity, and sensitivity to functioning difficulties in young children with ASD (Paul, Loomis, & Chawarska, 2011; Perry, Flanagan, Dunn Geier, & Freeman, 2009; Ray-Subramanian, Huai, & Ellis Weismer, 2011). The items of the VABS-II compose four domains: Communication, Daily Living Skills, Socialization and Motor Skills that are further condensed to an Adaptive Behavior Composite score (VABC). In study III, we applied this VABC summary score of the VABS to describe functional level of the participants (Table 1 and 2). Levels of adaptive functioning in ASD were generally lower than in NS.

In study IV, the summary VABC score was also chosen to operationalize impairment (or functioning) and as the most adequate proxy for the concept of impairment in DSM-5 as a global composite across contexts. There are several good reasons for using a global composite notwithstanding DSM-5 request for "clinically significant impairment in social, occupational, or other important areas of current functioning" and findings of uneven adaptive profiles in ASD. Firstly, DSM-5 endorses the use of a global impairment construct, for example by recommending the WHODAS [likewise DSM-IV recommended the global measures Global Assessment of Functioning [GAF] (APA, 1994, 2000) and Children's Global Assessment Scale [C-GAS] (Shaffer et al., 1983)] or stating that symptoms ''limit or impair everyday functioning'' (APA, 2000, p. 53). Secondly, following the DSM-5 logic of the separation of symptomatology from its consequences, i.e. impairment/dysfunction, it seems reasonable to treat impairment as a global construct rather than a composition of domains, of which at least two (socialization and communication) seem to repeat ASD symptomatology criteria (Goldstein & Naglieri, 2009; Rapee, Bőgels, van der Sluis, Craske, & Ollendick, 2012; White, Smith, & Schry, 2014). Thirdly, the composite score in our sample adequately reflects the even adaptive domain profile found in many previous studies of young children with ASD reviewed below. Different thresholds have been applied to designate impairment when using normed instruments (Bernier, 2012; Bolduc et al., 2011; Gathje, Lewandowski, & Gordon, 2008; Goldstein & Naglieri, 2009; Majnemer et al., 2012;

Naglieri, 2009; Ness et al., 2012; Papazoglou, Jacobson, & Zabel, 2013). We defined impairment using the three most commonly proposed thresholds: mild impairment cutoff was defined as an VABC of 1 SD below the mean ($\leq$ 84 standard score), moderate impairment as 1.5 SDs below the mean ($\leq$ 77 standard score), and severe impairment as 2 ($\leq$ 69 standard score) SDs below the mean.

In ASD, previous research has consistently indicated significant deficits in adaptive functioning already in early childhood (Harris, Handleman, Belchic, & Glasberg, 1995; Nah, Young, Brewer, & Berlingeri, 2014; Park, Yelland, Taffe, & Gray, 2012; Paul et al., 2011; Perry et al., 2009; Ray-Subramanian et al., 2011; Sutera et al., 2007). The same is even truer for older children, adolescents and adults (Bölte & Poustka, 2002; Fenton et al., 2003; Klin et al., 2007; Kraijer, 2000 for a a review of older studies; Liss et al., 2001; Paul et al., 2004; Szatmari, Bryson, Boyle, Streiner, & Duku, 2003; Tomanik, Pearson, Loveland, Lane, & Bryant Shaw, 2007). In general, adaptive functioning is comparably lower than IQ in individuals with average IQ but higher in individuals with below average IQ, and lower compared to IQ- and age-matched individuals (Bölte & Poustka, 2002; Klin et al., 2007; Perry et al., 2009). On a group level, individuals with ASD often exhibit an adaptive profile with the lowest scores on Socialization followed by Communication, Daily Livings Skills and Motor Skills (Klin et al., 2007; Kraijer, 2000; Perry et al., 2009; Saulnier & Klin, 2007). However, this pattern has not been identified consistently among younger children where the reported domain profile has been found to be rather even and therefore better reflected by the composite score (Fulton, Eapen, Crncec, Walter, & Rogers, 2014; Paul et al., 2004; Perry et al., 2009; Ray-Subramanian et al., 2011; Sutera et al., 2007). In previous studies, chronological age has commonly been substantially negatively associated with the level of adaptive functioning (Kanne et al., 2011; Klin, Saulnier, Chawarska, & Volkmar, 2008; Klin et al., 2007; Szatmari et al., 2003).

3.4 STATISTICS

In studies I and II, the interrater reliability of items and totals was analyzed using the *G(q,k)* estimator, calculated with the SAS macro G(q,k) provided by Putka et al. (2008). As the design in both studies was neither fully crossed nor nested (i.e. Ill-Structured Measurement Design [ISMD]) (Putka, Le, McCloy, & Diaz, 2008), the necessary assumptions for the most commonly used statistical methods like κ_w for multiple users and ICC were not fulfilled or their applicability incompletely described for the current design (Conger, 1980; Hallgren, 2012; Putka et al., 2008). The *G(q,k)* estimator used here estimates the rater main effect separately from the rater-subject interaction in unbalanced designs (Cicchetti & Sparrow, 1981; Putka et al., 2008) to produce coefficients comparable to ICC, and has been described as a modified ICC (1, k) (Shavelson & Webb, 2006). In ISMDs, it has been demonstrated that the *G(q,k)* estimator generates a more correct estimate of interrater reliability than ICC, and specifically to prevent from the risk of underestimating interrater reliability (Cicchetti & Sparrow, 1981; Putka et al., 2008). Like previous studies on the ADI-R (Le Couteur et al., 1989; Lord et al., 1994; Poustka et al., 1996; Tsuchiya et al., 2013) and the ADOS (Lord et

al., 1999; Lord et al., 2012), scores of 3 were converted to 2 in the analyses except for item A1 of the ADOS ("Overall Level of Non-Echoed spoken Language") where the scores of 0 to 4 were kept. Scores indicating "not applicable" (for example 7 and 8) were treated as missing values. Items that had fewer than three ratings other than zero were excluded from the analyses. Fleiss' κ for multiple raters (Fleiss, 1971) and Cohen's κ (Cohen, 1960) were used to analyze the interrater reliability for the diagnostic classification. PA, i.e. the number of agreements divided by the total number of observations, was used to estimate the interrater agreement for items and diagnostic classification (de Vet, Terwee, Knol, & Bouter, 2006). For the interpretation of the clinical significance of the interrater reliability coefficients, coefficients below .40 were considered as poor, .40–.59 as fair, .60-.74 as good and above .75 as excellent (Cicchetti, 1994). For PA, 70-79% agreement was evaluated to be fair, 80-89% good and above 90% excellent (Cicchetti, Volkmar, Klin, & Showalter, 1995).

In study III, diagnostic validity was determined by calculating sensitivity and specificity with 95% confidence intervals [CI] using the Wilson score method (Newcombe, 1998). Moreover, classification accuracy (% correctly diagnosed) and positive likelihood ratios [LR+] for different combinations of single and combined use of the ADI-R and the ADOS compared to clinical consensus diagnosis were calculated. The sample was divided into the developmental groups according to the new ADI-R research algorithms (NV, SW, and PH) when analyzing ADI-R data, ADI-R/ADOS combined, and according to the different ADOS algorithms (module 1 few to no words, module 1 some words and module 2 younger than 5 years) when analyzing the ADOS data alone in order to facilitate comparison with previous research. No separate analyses for the toddler module ($n = 5$) and module 3 ($n = 2$) were conducted. To examine the discriminative properties of the instruments, receiver operating characteristics [ROC] statistics with area under the curve [AUC] were computed for all developmental groups of the ADI-R and algorithms of the ADOS. To assess the agreement between the two instruments, Pearson's correlation coefficients between domain and total scores of the ADI-R and the ADOS were calculated. Fischer's Z transformation was used to analyze correlation coefficients differences (Cohen & Cohen, 1983). Categorical agreement between the ADI-R, ADOS, and clinical consensus diagnosis was determined by κ. Characteristics of correctly and misclassified children as well as the differences of total scores on the ADI-R and the ADOS between groups were analyzed with Bonferroni post hoc tests. Cohen's d was used to calculate effect sizes.

In study IV, percentages were calculated to examine the proportion of children with a DSM-IV-TR diagnosis that would fulfill or not fulfill the DSM-5 impairment criterion. The proportions of children with a DSM-IV-TR defined autistic disorder or PDD-NOS as well as with an incomplete DSM-5 ASD symptomatology according to our definition were compared with χ^2-statistics. The association of the different specifiers (symptom severity, intellectual and language impairment) and other variables (age and gender) with each level of impairment status, was analyzed with multiple logistic regression, adjusting for all the specifiers/factors at the same time, i.e. entered in a block in a single step. An α level of .05 was adopted (2-sided for the χ^2-analyses) for all analyses.

3.5 ETHICAL CONSIDERATIONS

All studies were approved by the Regional Board of Ethical Vetting, Stockholm, and for studies I-II, all caregivers had provided written consent of participation. All participants in the objectivity, diagnostic validity and impairment criterion studies have either participated in the studies as a part of a standard clinical evaluation at an outpatient clinic or within the framework of another research project. In no case, the testing, observations or interviewing has been carried out as an additional activity for any of the participants. Only anonymous data were used in the analyses. Therefore, there has been no direct advantage of participating in the studies but minimal risk as well.

4 RESULTS

4.1 STUDY I

The interrater reliability and agreement of all the individual algorithm items per domain were

- RSI: $G(q,k)_{median}$ = .98, range: .90-1.00, PA_{median} = 83%, range: 63-96%
- COM: $G(q,k)_{median}$ = .97, range: .91-.98, PA_{median} = 85%, range: 69-100%
- RRB: $G(q,k)_{median}$ = .94, range: .88-.96, PA_{median} = 79.5%, range: 71-100%
- the onset items: $G(q, k)_{median}$ = .96, range: .86-1.00.

The interrater reliability of the diagnostic criteria operationalizing ICD-10/DSM-IV criteria and ADI-R domains were

- RSI criteria ranged from $G(q,k)$ = .96 to .99, RSI domain: $G(q,k)$ = .99
- COM criteria ranged from $G(q,k)$ = .96 to 1.00; COM domain: $G(q,k)$ = .98
- RRB criteria ranged from $G(q,k)$ = .91 to .97; RRB domain: $G(q,k)$ = .97
- Onset criteria: $G(q,k)$ = .95.

The interrater reliability and agreement of diagnostic classification, i.e. both for each domain separately and the overall diagnostic classification were

- RSI: κ_{Fleiss} = .87, κ_{Cohen} = .93, and PA = 94%
- COM: κ_{Fleiss} = .88, κ_{Cohen} = .94, and PA = 96%
- RRB: κ_{Fleiss} = .50, κ_{Cohen} = .76, and PA = 84%
- onset: κ_{Fleiss} = .65, κ_{Cohen} = .83, and PA = 86%
- overall: κ_{Fleiss} = .68, κ_{Cohen} = .83, and PA = 84%.

4.2 STUDY II

The interrater reliability and agreement of all the individual algorithm items were

- module 1: $G(q,k)_{median}$ = .83, range: .23-.94, and PA_{median} = 60 %, range: 42-99%
- module 2: $G(q,k)_{median}$ = .74, range: .38-.91 and PA_{median} = 65%, range: 40-80%
- module 3: $G(q,k)_{median}$ = .74, range: .30-.89, and PA_{median} = 61.5%, range: 40-90%
- module 4: $G(q,k)_{median}$ = .75, range: .29-.92, and PA_{median} was 59.5%, range: 49-84%.

The interrater reliability of domains (SA and RRB) and overall totals of the ADOS-2 algorithms and the revised algorithm of module 4 were

- module 1: $G(q,k)$ = .91 for SA, $G(q,k)$ = .76 for RRB, and $G(q,k)$ = .92 for the overall total
- module 2: $G(q,k)$ = .86 for SA; $G(q,k)$ =.90 for RRB; $G(q,k)$ = .89 for the overall total
- module 3: $G(q,k)$ = .86 for SA; $G(q,k)$ = .45 for RRB; $G(q,k)$ = .85 for the overall total

- module 4: $G(q,k) = .92$ for SA, $G(q,k) = .73$ for RRB; $G(q,k) = .90$ for the overall total.

The interrater reliability and agreement of diagnostic classification, i.e. whether raters were consistent if the autism spectrum cutoff was met or not, were

module 1: $\kappa_{Fleiss} = .39$, $\kappa_{Cohen} = .71$, and PA = 82%

module 2: $\kappa_{Fleiss} = .22$, $\kappa_{Cohen} = .61$, and PA = 64%

module 3: $\kappa_{Fleiss} = .19$, $\kappa_{Cohen} = .62$, and PA = 74%

module 4: $\kappa_{Fleiss} = .55$, $\kappa_{Cohen} = .76$, and PA = 78%

modules 1-4 together: $\kappa_{Fleiss} = .38$, $\kappa_{Cohen} = .69$, and PA = 74.5%; excluding all participants with ADHD and PDD-NOS: $\kappa_{Fleiss} = .45$, $\kappa_{Cohen} = .75$, and PA = 86%.

4.3 STUDY III

Key findings in regards to diagnostic validity of the ADI-R and the ADOS separately and in combination for children 18-47 months old were

The ADI-R alone

- ADI-R clinical cutoff for all participants: sensitivity = 60% (range for the NV, SW and PH algorithms: 53-70%) and specificity = 76% (range: 69–81%). LR+ = 2.5 (range: 2.2–2.8)
- ADI-R research cutoff: sensitivity = 47% (range: 44-52%) and specificity = 94% (range: 91-96%). LR+ = 7.2 (range: 4.6-10.5)
- ADI-R adjusted cutoff: inspection of ROC-curves revealed that a lower cutoff yielded a more balanced sensitivity = 80% (range: 77–82%) and specificity = 61% (range: 60-62%). LR+ = 1.9 (range: 1.9-2.2)
- ADI-R AUC ranged between .74 and .79
- agreement between ADI-R clinical cutoff and clinical consensus diagnosis: $\kappa = .21$-.37; research cutoff and clinical consensus diagnosis: $\kappa = .25$-.40; and for the adjusted cutoff and clinical consensus diagnosis: $\kappa = .36$-.45
- age and VABC score associated with misclassification by the ADI-R: false positives [FP] (NS children falsely classified as ASD) were younger than true negatives [TN] (NS children correctly classified as NS) on the ADI-R ($F(3, 250) = 8.22$, post hoc: TN > FP, $p < .001$, $d = .90$), while false negatives [FN] (children with ASD falsely classified as NS) had higher Vineland-II scores than the true positives [TP] (children with ASD correctly classified) ($F(3, 249) = 34.27$, post hoc: FN > TP, $p < .001$, $d = .83$).

The ADOS alone

- ADOS autism spectrum cutoff for all participants: sensitivity = 97% (range for module 1 no to few words, module 1 some words and module 2 younger than 5 years: 94-100%) and specificity = 65% (range: 52–76%). LR+ = 2.8 (range: 2.1-3.9)
- ADOS autism cutoff for all participants: sensitivity = 84% (range: 81-94%) and specificity = 82% (range: 81-83%). LR+ = 4.6 (range: 4.4-5.4)
- ADOS AUC ranged between .91 and .95
- agreement between ADOS autism spectrum cutoff and clinical consensus diagnosis for different modules/algorithms: κ = .60-.72; ADOS autism cutoff and clinical consensus diagnosis: κ = .57-.75
- no significant associations between any of the participant characteristics and misclassified children (FP and FN).

The ADI-R and the ADOS combined

- the different combinations of ADI-R (clinical and research cutoffs) and the ADOS (autism spectrum and autism cutoffs) for all participants yielded sensitivities of 42-58% and specificities of 92-99%. LR+s = 7.4-41.0
- the ADI-R adjusted cutoff and the autism spectrum cutoff for the ADOS for all participants gave a sensitivity = 78% and specificity = 88%. LR+ = 6.2; the ADI-R adjusted cutoff and the autism cutoff for the ADOS for all participants gave a sensitivity = 68% and specificity = 90%. LR+ = 6.6
- agreement between the ADI-R clinical and research cutoffs and the ADOS autism spectrum cutoff for all participants: κ = .23; ADI-R clinical and research cutoffs and the ADOS autism cutoff: κ = .29 and κ = .30 respectively
- agreement between the adjusted ADI-R cutoff and the ADOS autism spectrum and autism cutoffs for all participants: κ = .31 and .34
- correlations between the ADI-R and the ADOS total scores: ADI-R NV developmental cell: r = .53 (p < .001); ADI-R SW developmental cell: r = .31 (p = .004); and ADI-R PH developmental cell: r = .42 (p < .001). The correlations between the SA domain of both instruments showed the same pattern: NV r = .50 (p < .001), SW r = .28 (p = .009), and PH r = .45 (p < .001), while the RRB domains correlated weakly for the SW and PH groups: r = .19 (p = .076 and p = .062) but stronger for the NV group (r = .40, p = .001).

4.4 STUDY IV

Key findings in regards to fulfilling the DSM-5 impairment criterion and factors associated with the fulfillment were

- application of the cutoff for the mild impairment level (i.e. VABS-II ABC = 1 SD below the mean): 88 % or 112 of the 127 children fulfilled the impairment criterion

(proportionally more children with an incomplete DSM-5 symptomatology documented on the ADOS-2 and the ADI-R in the non-impaired group: $\chi^2(1) = 6.97$, p = .008)

- application of the moderate cutoff (= 1.5 SDs below the mean): 69 % or 88 children fulfilled the impairment criterion (the proportion of children with an incomplete DSM-5 symptomatology did not differ significantly between the non-impaired and the impaired groups: $\chi^2(1) = 3.22$, p = .073)
- application of the severe impairment cutoff (2 SDs below the mean): 33 % or 42 children fulfilled the criterion (the proportion of children with an incomplete symptomatology did not differ between groups: ($\chi^2(1) = 3.13$, p = .077)
- DSM-5 ASD specifiers and variables (intellectual and language impairment, symptom severity, age and gender) associated with impairment status when applying the mild impairment cutoff, none; using the moderate impairment cutoff: intellectual impairment (Odds ratio [OR] = 2.92, 95 % CI 1.05-8.12); using the severe impairment cutoff: intellectual impairment (OR = 8.82, 95 % CI 3.20-24.33), language impairment (OR = 7.10, 95 % CI 2.02–24.98), and being a girl (OR = 3.33, 95 % CI 1.04-10.63).

5 DISCUSSION

5.1 STUDY I

In study I, the objectivity, operationalized as interrater reliability, of the ADI-R was examined for the first time in a naturalistic multicenter clinical setting as well as in Scandinavia. Consequently, the results of the current study add information to previous findings on the interrater reliability of the ADI-R in terms of generalizability to clinical practice of ASD and to the cross-cultural transferability of the instrument's objectivity.

The interrater reliability on item level was comparable to previous studies conducted in research setting, exceeding $G(q,k)_{median}$ values of .90 for all diagnostic domains, while somewhat lower $PA_{medians}$ compared to earlier reports. On ICD-10/DSM-IV criteria level, the interrater reliability of the ADI-R was also excellent with $G(q,k)$ estimates above .90 for all criteria, thus similar to previous studies from research settings.

Besides this study and the previous ones of the ADI-R and the DSM-IV field trials (Klin, Lang, Cicchetti, & Volkmar, 2000) reporting on the interrater reliability or objectivity of the assessment of ASD criteria, there is a scarcity of published data despite the existence of several other diagnostic interviews that have gained popularity. One example with published interrater reliability data though is the DISCO, a standardized interview that compared to the ADI-R also covers more general developmental issues as well as assesses the needs of the patient. The DISCO also comprises over 300 items and 500 related ratings, making it even more time-consuming than the ADI-R to administer. Promising interrater reliability on item level but with a different analytic approach than the ADI-R studies have been reported in a couple of studies including one from Scandinavia (61.6% of all items κ/ICC ≥ .75, Nygren et al., 2009; 80% of all items κ/ICC ≥ .75, Wing et al., 2002). On the ICD-10 diagnostic criteria, only one study has reported interrater reliability estimates (Leekam, Libby, Wing, Gould, & Taylor, 2002): RSI criteria ranged from $\kappa = .60$ to ICC = .96, COM between $\kappa = .77$ and ICC = .95, RRB between ICC = .84 and .93 and ICC = .82 for onset. Hence, this does not indicate that the DISCO's considerably longer diagnostic algorithm for autism (89 or 91 items depending on the version), in relation to that of the ADI-R, produces higher interrater reliability. The computer-based 3di is another example of a diagnostic interview operationalizing the ICD-10/DSM-IV criteria and that has an even longer ADI-R related diagnostic algorithm (122 items). Data on interrater reliability has only been reported on domain level (ICC = .85-1 for ASD and ICC = .81-1 for non ASD) (Skuse et al., 2004). The diagnostic interview PIA is respondent-based and do, probably because of that, not report any interrater reliability (Stone, Coonrod, Pozdol, & Turner, 2003; Stone & Hogan, 1993) whereas the interrater reliability for the dichotomously assessed diagnostic criteria of the ASDI, a diagnostic interview operationalizing the Gillberg and Gillberg criteria of Asperger's Disorder (I. C. Gillberg & Gillberg, 1989), was $\kappa = .91$ for all the 20 items together (C. Gillberg et al., 2001). As regards the ASD sections of the broadband diagnostic interviews for children and adolescents like the DAWBA (Goodman et al., 2000) there are no interrater

reliability data at all available while for the K-SADS-PL (Kaufman et al., 2009; Kaufman et al., 1997), one study has reported data on diagnostic classification within the autism spectrum (differentiating between autistic and Asperger's disorder as well as PDD-NOS) (ICC = .72-.85 for present-diagnoses and ICC = .41-1 for ever-diagnoses (Zavaleta-Ramirez, Nafate-Lopez, Villarreal-Valdes, Ulloa-Flores, & Albores-Gallo, 2014). From the summary of the results from previous research on the ADI-R, especially when adding the results from the current study, and the other diagnostic interviews reviewed here, the ADI-R appears as both the most well-documented and best performing diagnostic interview to date for an objective assessment of ASD criteria in clinical practice and research.

5.2 STUDY II

In study II, the objectivity, operationalized as interrater reliability, of the ADOS was examined for the first time in a naturalistic multicenter clinical setting as well as in Scandinavia. Overall, our results contribute to a better understanding of to what extent previous psychometric findings of the ADOS from research settings could be generalized to daily clinical practice where the users of the instrument typically are only basically trained on the ADOS, not "research reliable", i.e. raters thoroughly prepared and continuously and systematically calibrated. As the ADOS is increasingly used and relied upon for diagnostic decision-making around the world in clinical practice as well as recommended by different national guidelines and professional societies, this is useful information.

The interrater reliability for items and domain totals was basically in the same good to excellent range as reported in previous studies for extensively trained and calibrated research reliable raters, although the PA for items was substantially lower. There was no consistent pattern of items with low and high interrater reliability across modules, except for a possible tendency that the ratings of certain RRB items like sensory interests, mannerisms, repetitive interests and compulsions and rituals to be less reliable than those of the SA items, like in previous studies. The interrater reliability of ADOS-2 classification (autism spectrum versus non spectrum) was in the acceptable range but lower than in previous studies. Only module 1 scored in the range of good interrater reliability using PA, modules 3 and 4 in the fair range, and module 2 in the poor range; the interrater reliability of all modules taken together was fair. This study compared to the previous ones differs in regards to a number of aspects that might have contributed to the lower rate of interrater reliability of ADOS-2 classification. To begin with, our spontaneous and naturalistic multicenter design with mostly not research reliable raters contrasts with previous research with highly trained and experienced raters following a systematic and continuous calibration activity and might have influenced the convergence of classification. Furthermore, the composition of the sample in the present study might have influenced to our results. Our sample was characterized by a large proportion of participants with PDD-NOS, i.e. a lesser variant of ASD, and ADHD, while the Lord et al. (Lord et al., 2000; Lord et al., 2012) and Bölte and Poustka (Bölte & Poustka, 2004) studies were dominated by individuals with core autistic disorder. When Lord et al. (Lord et al., 2000; Lord et al., 1999) included participants with PDD-NOS, the interrater

reliability in terms of PA decreased, even though still in the good to excellent range (81-93%). Analogously, excluding participants with PDD-NOS and ADHD in the present study resulted in increased interrater reliability of classification from a fair (74.5%) to a good level (86%). Hence, this suggests that samples characterized by larger diagnostic heterogeneity and more mild cases of ASD, i.e. resembling clinical practice, makes the ADOS ratings and diagnostic ADOS-2 classification less objective than compared to more homogenous samples of ASD (like in the studies from research settings), at least with non-expert raters.

5.2.1 Limitations, studies I and II

Both studies I and II suffers from several limitations. First, it would had been beneficial with larger and more diverse samples in both studies as they used the same complex unbalanced design to get the opportunity to examine for example how certain participant characteristics and diagnoses might influence the different levels of the objectivity or interrater reliability of the ADI-R and the ADOS. Second, none of the studies examined diagnostic validity but solely objectivity. Diagnostic validity is dependent on objectivity, but high levels of objectivity does not automatically translate into other properties such as diagnostic validity, and therefore only analyzing the objectivity leaves us with incomplete psychometrics regarding the ADI-R and the ADOS in clinical settings. However, as these studies are the first to examine objectivity in clinical practice among interviewers and examiners with varying experiences and training in using the ADI-R and the ADOS, they add new information of the instruments' value in clinical practice and demonstrate that some crucial prerequisites for diagnostic validity were fulfilled.

5.3 STUDY III

In study III, the psychometric properties were examined of the separate and combined use of the ADI-R and the ADOS applying their new diagnostic algorithms to children below 4 years of age in a clinical setting. Few studies have done this before and no study in Scandinavia. Our results indicate that the diagnostic validity of the combined use of the ADI-R and the ADOS in children of this age range with unclear developmental concerns is superior to separate use of each instrument. The ADOS autism cutoff alone performed almost at the same level. The diagnostic validity was comparable to the results from US research settings for both instruments, separate and combined, but for the ADI-R only after cutoffs adjustment.

For the ADI-R alone sensitivities were overall substantially lower than specificities, which contrasts the two previous studies examining the new ADI-R research algorithms in mixed research and clinical samples, where higher and more balanced sensitivities and specificities were found (Kim et al., 2013; Kim and Lord, 2012b). The adjusted, i.e. lower, cutoffs yielded better balanced sensitivities and specificities, and basically eliminated the differences with the US samples as well as increased the classification accuracy. Nonetheless, LRs+ that were lower than in previous studies, decreased with the adjusted cutoffs, and this together with that none of the AUCs exceeded a fair level indicate that the ADI-R had limited classification accuracy in our clinical sample of toddlers and young preschoolers with serious

developmental concerns, and that the ADI-R did not classify children as efficiently as in the US samples.

Overall the sample characteristics and how they were associated with domain totals in this study were similar to previous research with some exceptions that might have accounted for some of the differences of accuracy between our study and the US ones. First, all domain totals in our study were lower and in particular those of the RRB domain than in the three US samples (Kim et al., 2013; Kim and Lord, 2012b). As these lower levels of ADI-R totals were not associated with any differences of child characteristics or corresponding lower levels of ADOS totals we speculate that the differences might be linked to parent/caregiver characteristics, to the fact that we examined a purely clinical sample, as well as cultural factors. We have experienced that parents of young children in clinical practice often are either reluctant to describe or unaware of abnormality or developmental alteration of their child's behavior. Moreover, cultural differences between Europe and the US have been found in other parent-based instruments assessing autistic symptomatology (see for example, Bölte, 2012; Bölte, Poustka, & Constantino, 2008) as well as in a recent European multicountry study of the new research algorithms of the ADI-R (de Bildt, Sytema, Zander, et al., 2015). Furthermore, some of the differences of domain totals level might also have been due to chance rather than being significant differences as there is an overlap in confidence interval between this and at least one of the US samples in many instances. Second, age was associated to misclassification of children the opposite way in our study where NS children misclassified as ASD were younger than the correctly classified children with NS compared to that of the Kim and Lord study (S. H. Kim & Lord, 2012a). This indicates that age effects might still be an issue even in the new research algorithms of the ADI-R, producing a certain degree of overinclusivity in young children.

The evaluation of the ADOS in this study differed from previous research in the following ways: no other study has presented data for children up to 47 months only from purely clinical practice, separately for each of the revised algorithms and classified the children in an ASD versus NS group. Previous research has either included children up to 12 years for the algorithms of module 1 and 2 used in this study (de Bildt et al., 2009; Gotham et al., 2007, 2008; Oosterling et al., 2010), reported data from at least partly research settings (the majority of the reviewed studies), lumping the results from all algorithms together if reporting from the same restricted age range as this study (e.g., S. H. Kim & Lord, 2012a), or analyzed diagnostic validity separately for autistic disorder excluding other ASDs versus NS on one hand and nonautistic disorder ASDs versus NS on the other. The participants in this study had higher NVIQ than those in the two Gotham et al. (2007; 2008) studies, the Gray et al. (2008), and the De Bildt et al. (2009) studies but were lower functioning than in Oosterling et al. (2010). However, all these studies reported data for children with autistic disorder and ASD apart and also for children up 12 years of age. Therefore, a direct comparison with other studies is somewhat compromised but especially when taking the variation of the reported domain totals into consideration, it is clear that our ASD sample did not differ from the reviewed studies in any crucial way, and especially not from the point of

view of diagnostic validity of the ADOS.

In our study, the diagnostic validity of the ADOS autism spectrum cutoff showed excellent sensitivities together with lower specificities like previous research by Kim and Lord (2012a), Molloy et al. (2011) (except for module 2), and Risi et al. (2006) to which the results of Oosterling et al. (2010) contrasted with lower sensitivities and higher specificities for module 1 (some words) and module 2. The autism cutoff yielded more balanced sensitivities and specificities ($\geq 81\%$), but considerably lower than those reported by Gotham et al. when using the autism cutoff for children with core autism (not other ASDs) versus NS (2007; 2008). However, compared to the sensitivities and specificities produced by the autism spectrum cutoff in the Gotham et al. studies as well when taking the 95% confidence intervals into account (when such are reported), most differences were attenuated, disappeared or were reversed between this and previous studies. LR+s were modest in our study but comparable with previous research (S. H. Kim & Lord, 2012a). AUCs in our sample were excellent for all algorithms.

In our sample, the best diagnostic validity was produced by combining the ADI-R, and the ADOS, using the adjusted ADI-R cutoff and the ADOS autism spectrum cutoff, yielding sensitivities and specificities substantially in the same range as those in Kim and Lord (2012a) and Risi et al. (2006: using the standard algorithms with children below 36 months of age), except for SW group in our study that had a somewhat lower sensitivity. Used separately, the ADOS ($\kappa = .57–.75$) was more accurate than the ADI-R ($\kappa = .27–.45$). The agreement of the ADI-R and the ADOS expressed by correlations of domain totals was limited but basically comparable to that of the Kim and Lord study (2012a). Older studies using the standard algorithms of both instruments, Risi et al. (2006) and Le Couteur et al. (2008) have reported higher correlations between domain totals while Ventola et al. (2006) found a weak agreement measured by kappa statistics between the ADI-R and the ADOS classifications.

5.3.1 Limitations, study III

Study III suffers from some notable limitations. First, the clinical consensus diagnosis was not completely independent of the ADI-R and the ADOS as their results were part of the information used in the diagnostic decision-making. This is consistent with most other diagnostic validation studies (Gotham et al., 2007, 2009; S. H. Kim & Lord, 2012a, 2012b; S. H. Kim et al., 2013; Le Couteur et al., 2008; Risi et al., 2006). However, the current study also employed a variety of other sources of information in the diagnostic decision-making supplementing, and balancing the results of the ADI-R and the ADOS, in particular a prolonged naturalistic observation of the child in his/her preschool. Second, $n = 30$ children of the NS group did not get any DSM-IV-TR diagnosis after evaluation. Introducing typically developing children in a study of diagnostic validity is likely to overvalue especially the specificity (see, for example, Kim and Lord, 2012b: where this happened). However, these children were clinically referred and assessed children because of experienced developmental

concerns, therefore in most cases not typically developing in a narrow sense. Omitting them from the analysis actually led to a slightly decreased specificity.

5.4   STUDY IV

In study IV, the potential effect of new DSM-5 impairment criterion for ASD on diagnosis rates in toddlers and young preschoolers was examined for the first time in addition to the impact of the DSM-5 specifiers symptom severity and intellectual and language impairment as well as age and gender on impairment status. Despite the abundance of studies on the effect of the new DSM-5 ASD symptomatology criteria on diagnosis rates indicating that a considerable part of the children with an ASD diagnosis according to DSM-IV-TR would not fulfill the new DSM-5 criteria, remarkably few if any studies have examined the effect of the DSM-5 impairment criterion on ASD diagnosis. Our results indicate that the application of the new impairment criterion might constitute a so far largely overlooked problem for diagnosing ASD in infants, toddlers and young children compared to DSM-IV-TR. As early diagnosis is a prioritized objective in clinical practice of ASD and an increasing number of children with ASD are getting an early diagnosis, this matter warrant more attention. The introduction of the impairment criterion might also be of a far greater importance for the youngest children with ASD than the new DSM-5 symptomatology criteria.

The effect of the new impairment criterion on diagnosis rates was, depending on the threshold chosen (1 SD, 1.5 SDs or 2 SDs below average on the VABS-II) that between 12 and 67 % of the children below 47 months of age fulfilling DSM-IV-TR criteria for ASD did not fulfill the new DSM-5 impairment criterion while it was only 9 % of the participants that did not fulfill the new DSM-5 symptomatology criteria for ASD. In regards to the impact of the ASD specifiers, there was no significant association between symptom severity and impairment status. Intellectual impairment, on the other hand was the specifier the most strongly associated with impairment status, like in previous studies (Ray-Subramanian et al., 2011), but only when applying the moderate and severe cutoffs, not the mild impairment cutoff. In addition, concerning the severe impairment cutoff, language impairment was also associated to impairment status, as well as female gender. Age was not significantly associated to impairment like in previous research, which is probably due to our sample's restricted age range compared to earlier studies. Thus, the results of the current study suggest that functional impairment in toddlers and young children is more strongly linked to the presence of ASD non-specific problems or specifiers rather than to symptom severity.

5.4.1   Limitations, study IV

Study IV suffers from several limitations. First, the relatively small sample of young children might have affected the generalizability of the present findings. For example, the confidence intervals of the analyses of associations between ASD specifiers and other factors with impairment were quite large. Moreover, due to the restricted age range in the current sample, conclusions about the validity of the results for older individuals are mere speculation. Second, the measures of impairment and symptom severity applied in the current study might

have driven the results. Although the DSM-5 endorses an ICF-based view concerning impairment, both ''significant impairment'' and how it is manifested in the ''social, occupational, or other important areas of current functioning'' at different ages and in various contexts are still to be defined. However, the VABS-II seems currently to be one of the best documented measures in this area, and closely resembling the unpublished Child WHODAS (Canino, 2013), which is ICF derived. Other conclusions might have been drawn if using other conceptualizations of impairment and/or other instruments or drawing on information from other settings, for example the child's preschool. In regards to symptom severity, the ADOS-2 CS is in our opinion currently the most appropriate and perhaps also the only measure available for quantifying the ASD symptomatology independently of age and language level.

5.5   GENERAL DISCUSSION

An important motivation to use standardized diagnostic instruments in clinical practice and research is their potential ability to improve the objectivity of the assessment of symptoms, diagnostic criteria and diagnostic classification. The objectivity, i.e. to what extent a test result is independent of the examiner, often operationalized as interrater reliability, is a crucial but not the only important psychometric aspect of a diagnostic instrument. However, it is an especially valued property of a diagnostic instrument in psychiatric practice. This is mainly due to the fact that diagnostic categorization often are dependent on clinicians ability to capture the relevant information through directly observing behaviors or listening to descriptions of behavioral symptoms, and that the behavioral definitions or criteria, operationalized by these instruments, are the most concrete, and sometimes only expression of the disorder we have.

From the point of view of objectivity, previous studies from research settings and our studies from clinical practice have demonstrated the psychometrical soundness of the ADI-R and the ADOS for sampling separate symptoms and criteria of ASD as well as for diagnostic classification. We demonstrated that the excellent levels of objectivity of the ADI-R reported from research settings could be reproduced in clinical practice and to a high degree, even though lower than for the ADI-R, for the ADOS as well. Indeed, this suggest that the use of the ADI-R and the ADOS represents more objective means of collecting diagnostically relevant information than for example using the DSM criteria alone to organize clinical impression, like in the field trials of DSM-IV and DSM-5, other forms of history taking or compared to other methods used in psychiatry and medicine in general. In the field trials of DSM-IV, the objectivity estimates for the 12 different criteria of autistic disorder ranged from $\kappa = .58$ to .79 (Klin et al., 2000), to be compared to $G(q,k) = .91$-1.00 for the operationalized DSM-IV diagnostic criteria of the ADI-R and $G(q,k)_{medians}$ between .74 and .83 (range for all items of all modules: $G(q,k) = .23$-$.94$) for single symptoms/items of the ADOS in the current studies. The interrater reliability of classification, using diagnostic criteria only, of autistic disorder (all other ASDs excluded) versus a non-spectrum disorders according to DSM-IV was $\kappa = .95$ (Volkmar et al., 1994) and ASD versus non ASD according to DSM-5 $\kappa = .69$

(Freedman et al., 2013) in the respective field trials while the correspondent estimates in the current study for the ADI-R was $\kappa = .83$ and $\kappa = .69$ and 75 (when excluding participants with PDD-NOS and ADHD) for the ADOS. Moreover, data from the field trials also indicate that clinician's experience and the use of DSM-IV criteria or not when assigning diagnosis influence the objectivity estimates for autistic disorder: the interrater reliability for experienced versus inexperienced raters not using DSM-IV criteria was $\kappa = .84$ and $\kappa = .34$ while the corresponding estimates were $\kappa = .94$ and $\kappa = .59$ when DSM-IV criteria was used (Klin et al., 2000). Definitely, the objectivity estimates of classification of the ADI-R and the ADOS hold their own in comparison with those of the other instruments specifically designed to assess ASD reviewed earlier as well as other mental disorders (Dittmann, von Cranach, & Eckermann, 1990: $\kappa = .79$ for ICD-10 personality disorders; Flaum & Andreasen, 1995: $\kappa = .50$ for DSM-IV schizophrenia symptoms; Freedman et al., 2013: $\kappa = -.03-.78$ for DSM-5 diagnoses; Regier, Kaelber, Roper, Rae, & Sartorius, 1994: $\kappa = .52-.83$ for 1 to 4 character diagnostic categories of ICD-10; Sartorius, Üstün, Korten, Cooper, & van Drimmelen, 1995: $\kappa = .65$ for all categories of ICD-10) and for the objectivity of other assessment methods used in healthcare (Gullett et al., 2015: ICC = .615-.904 for ultrasound; Koran, 1975a, 1975b for a multitude of methods; Kushnirsky et al., 2015: ICC = .73-.79 for detection of brain metastasis with MRI; Olson et al., 2015: $\kappa = .40-.62$ for pupillometry; Overbury, Murtaugh, Fischer, & Frech, 2015: $\kappa = .50$ for capillaroscopy; Quigg et al., 2015: $\kappa = .57$ for electrocorticograpical recordings).

Although objectivity or interrater reliability is a crucial element of an instrument's psychometric properties indicating successful standardization of data collection, scoring and interpretation, objectivity together with different aspects of reliability as a whole are only prerequisites for high validity. In general, previous studies have reported satisfactory diagnostic validity for both the ADI-R and the ADOS apart for children from 12 months of age up to adulthood. However, previous research (S. H. Kim & Lord, 2012a; Risi et al., 2006) and the current study have also demonstrated the added contribution of using *both* instruments in combination in diagnostic decision-making instead of only one of them, despite the fact that the diagnostic validity of the ADI-R was less satisfactory in comparison with that of the ADOS in the current study.

The ADI-R and the ADOS appear to cover different but equally important dimensions of the autistic symptomatology, for example demonstrated by the lack of very substantial correlations between the two instruments. The ADI-R, through descriptions by a caregiver, has the potential to contribute with unique, essential and reliable information regarding early development, trajectories and timing of symptom onset, and descriptions of behavior within a variety of contexts over time, retrospective information that is hardly impossible to obtain through any other means. In general, the ADI-R is a highly objective instrument although the diagnostic validity in the current study was lower than for the ADOS. The objectivity of the ADI-R in this particular sample was not assessed but we still must be aware of the possibility that a verbal exploration of a caregiver in the context of a diagnostic interview might be perfectly objective, in the sense that two raters would code the enunciations of the caregiver

44

in the same way, but must not for that sake be valid, in the sense that the enunciations reflect the child's true behavior or that these behaviors must reflect the presence of ASD. Therefore, the information of the ADI-R and other diagnostic interviews is limited without other types of information, especially an expert's direct observation like the ADOS. The ADOS, on the other hand, samples ASD symptoms during a very short interactive observational session, which also contributes with unique and essential information not possible to obtain by any other means. Due to its design, not all criteria and aspects of ASD are covered and there is a risk that the individual will not exhibit all his/her typical behavioral characteristics due to the brevity of the administration. Moreover, the quality of the information obtained from the ADOS compared to the ADI-R is more dependent on the skill of the examiner, which is reflected in the generally lower, but still satisfactory objectivity of the ADOS. The use of the ADOS in isolation for diagnostic decision-making must therefore be considered as limited as the single use of the ADI-R. Hence, for the diagnostic assessment of ASD in clinical practice and research, the first choice is the combined use of the ADI-R and the ADOS. Still, this does not mean that the results from the ADI-R or the ADOS neither separately nor in combination could supplant a best estimate diagnosis as gold standard from the standpoint of objectivity and diagnostic validity.

Objectivity and diagnostic validity are two separate aspects of the psychometric properties of a standardized diagnostic instrument. However, in psychiatry, objectivity is sometimes seen as "the first test of validity for a diagnosis" (Freedman et al., 2013) and if not as equivalent to validity so very close to it. In the current diagnostic classifications, psychiatric diagnoses are conceptually latent constructs, i.e. the true nature or pathophysiology of for example ASD is considered latent and cannot currently be directly observed and described scientifically, and therefore not used for diagnostics. The behavioral manifestations of the diagnosis, summarized in the diagnostic criteria, are thought of as indicators of these latent phenomena but not as the true disorder entity. The behavioral symptoms, though, are the most concrete and often only detectable manifestation of a mental disorder and have come to play an extremely important role in psychiatric practice, not the least because of their clinical utility. Furthermore, the objectivity or interrater reliability represents the most elaborated (and authoritative?) scientific foundation for proving the existence or presence of a psychiatric diagnosis through the assessment of its diagnostic criteria. Probably because of that, criteria have come to be treated as the disorder itself or reify the disorder, despite criteria's more limited role of being indicators of latent disease entities, guidelines for diagnostic assessment and heuristic tools only (Hyman, 2010). The classifications obtained from standardized diagnostic instruments like the ADI-R and the ADOS, operationalizing diagnostic criteria, have also tended to be used as reified disorders (Lord et al., 2014), for example when the results are used as a proxy for diagnosis in research (Bryson et al., 2007; Ozonoff et al., 2014; Wolff et al., 2012), or when clinicians, as we have experienced, heavily rely on the results from especially the ADOS in diagnostic decision-making. Consequently, when diagnostic criteria are used as the criterion against which a standardized diagnostic instrument –

operationalizing these same diagnostic criteria – is validated, this procedure resembles more of a test of objectivity than of validity.

Recently, in the mental health field, as we have already mentioned, the concept of impairment has gained greater importance as a dimension in its own right, separated from the defining psychopathology and more directly linked to functioning, need for support and intervention planning than the psychopathology. In DSM-5, impairment is often a required criterion for diagnosis, like in ASD, and conceptualized as caused by the psychopathological symptoms (that in turn are caused by the latent disease entities). However, the result from the current study as well as from previous research on adaptive functioning in ASD does not necessarily support the DSM-5 notion, at least not in the given age group of our study, that symptoms should cause impairment and be directly linked to the need and level of support in ASD. There are a number of issues in regards to the definition, operationalization and measurement of impairment as well as of symptom severity and the significance and role of the ASD specifiers in this context that need further exploration.

First, there is no consensus at what point a functional problem turns into impairment. The recommendation is to use normed instruments and applying specific norm-based cutoffs. Following this, impairment would be operationalized in relation to the norm, i.e. equivalent to a certain proportion of the population with the lowest scores. Using our mild impairment cutoff (1 SD below the mean), favored by Naglieri (2009), around 16 % of the general population would statistically fall in the impairment range. The application of such a low threshold seems controversial for a neurodevelopmental disorder such as ASD, affecting 1–2 % of the general population. Our moderate and severe cutoffs (1.5 and 2 SDs below the mean), corresponding to 7 and 2 % of the general population, respectively, seem spontaneously more realistic for this purpose. Though, under these circumstances, as our study have demonstrated, and could be extrapolated from previous research on adaptive functioning in ASD, a substantial proportion of toddlers, young preschoolers, and even older individuals fail to fulfill the impairment criterion (Gathje et al., 2008; Papazoglou et al., 2013). In DSM-5, despite the recommended use of normed instruments (i.e. WHODAS), impairment appears to be linked to the need for support in the first hand (see for example the severity specifier in ASD). Either way, the use of specific cutoffs or levels of need for support, the possibility for very young children to fulfill DSM-5 diagnostic criteria seems to be compromised. Some children would not fulfill this criterion even when using the very inclusive mild impairment cutoff, and employing need for support does not seem to change this. By definition, very young children in general need and should get a lot of care/service from their caregivers, which in many cases might complicate the assessment of need for ongoing support/impairment from typical age-related needs/functioning at this age. A negative consequence of a too strict application of the impairment criterion in the youngest children with ASD symptomatology could mean delayed diagnosis and early intervention preventing secondary prevention and causing avoidable burden and potentially worse outcomes.

In the current study, we defined impairment as deficits in adaptive functioning, a well-documented dimension in ASD. Through the recommended use of the WHODAS, DSM-5 endorses the conceptualization of impairment according to the ICF. The close relationship between the ICF components activities and participation and adaptive functioning has been demonstrated (Fabiano & Pelham Jr., 2009). Truly, the most elaborated framework for assessing impairment is the integrative biopsychosocial model of functioning, disability, and health of the ICF (WHO, 2001, 2007). The expectation is that a more specific and exhaustive model of functioning and improved measures of impairment, tailored for individuals with ASD in all ages will be available in the near future as a result from ongoing development of ICF core sets for ASD (Bölte et al., 2013; Gan, Tung, Yeh, & Wang, 2013). Especially important, the ICF model of functioning includes the component environmental factors, an aspect not explicit in the DSM-5 conceptualization of impairment and not included in the concept of adaptive functioning. Previous research has demonstrated that environmental factors are of utmost importance in order to characterize impairment/functioning properly in an individual, as these factors to a substantial degree determine individual outcomes (Goldstein & Naglieri, 2009; Rapee et al., 2012). Lastly, besides the ''objective'' characterization of the individual's impairment/level of functioning, the experienced quality of life, i.e. the subjective reality of the individual, also ought to be taken into account as an important dimension of mental disorders (Kuhlthau et al., 2010).

Finally, the definition, operationalization and measurement of ASD (symptom) severity need to be addressed. Typically, symptom severity is defined as number of symptoms and/or the intensity of symptoms (APA, 2013; Bernier, 2012). However, the symptom manifestation of ASD varies depending on age, IQ and language level. Therefore ASD severity ratings derived from counting symptoms and/or assessing their intensity are not absolute and need adjustment for third variables. Although the ADOS-2 and the ADI-R generate adjusted total scores from a symptom count including a form of intensity assessment, our and other studies indicate that these ratings do not necessarily translate into the same degree of impairing ''autistic'' problems in everyday life. The ADI-R, for example, adjusts for age and language level through specifically sampling symptoms from the period between 4 to 5 years for all individuals older than 4 years and applies different algorithms for verbal and nonverbal individuals. The ADOS has addressed this problem with sufficient success by using different modules and algorithms depending on language level and age, and specifically by the creation of the ADOS-2 Comparison Scores. Moreover, although the "raw" severity ratings of the ADI-R, i.e. the domain totals of the algorithms with higher scores indicating more symptoms and larger symptomatic intensity, the coding of the ADOS seems to differ somewhat from that logic. Due to its form and brevity even very subtle deficits and deviances are coded for diagnostic/discriminative purposes and these do not necessarily and/or always translates into real "autistic" problems of the same intensity in the reality (which is also underscored in the manual, Lord et al., 2012). Besides, none of these instruments have originally been designed to measure symptom severity, and the authors have also warranted against trying to do this, at least without taking IQ and age into account (Lord et al., 2012, p.

248; Rutter et al., 2003, p. 2). The same seems true for other instruments sampling ASD symptoms, like the Social Responsiveness Scale [SRS] (Hus & Lord, 2013). Thus, the issue here is, as we have objective measures of ASD symptomatology, how to design measures of symptom severity that account for age, language and developmental level (Hus, Bishop, Gotham, Huerta, & Lord, 2013) and then continue to explore how symptom severity is linked to the latent disease entities as well as affects functioning and impairment.

6 CONCLUSION AND FUTURE DIRECTIONS

6.1 CONCLUSIONS

The results of this thesis and previous research demonstrate that the ADI-R and the ADOS assess ASD more reliably than other similar instruments for ASD, other mental disorders and other methods. We showed that the ADI-R in regards to the diagnostic criteria of ASD according to ICD-10/DSM-IV produced assessments as objective as currently possible, even among a group of mainly not research reliable clinicians in clinical practice in Sweden. We also demonstrated that the ADOS in terms of objectivity is good enough to warrant the use of it as a psychometrically sufficiently sound method to support diagnostic decision-making even in clinical settings but this and other research do not favor its use as a stand-alone diagnostic decision-maker, especially not in clinical practice and in heterogeneous samples. In terms of diagnostic validity, the results from our study of young children demonstrated like previous research that the combined use of the ADI-R and the ADOS yielded better diagnostic validity for ICD-10/DSM-IV ASD than the single use of each instrument, and this despite the fact the ADI-R exhibited more limited diagnostic accuracy than previous research and the ADOS in our study. Overall, this endorses the universal use of the ADI-R and the ADOS in combination in everyday clinical practice and research for all age groups as first choice diagnostic instruments but for the ADOS in clinical practice it also underscores the importance of advanced training to reach research levels of objectivity. Moreover, in regards to the new diagnostic definition of ASD in DSM-5, we showed that applying the new impairment criterion strictly might compromise the possibility for very young children to get an ASD diagnosis despite manifesting the defining symptomatology, a previously unnoticed and not studied effect of the new diagnostic criteria. Finally, our findings in this group of young children indicate that there are substantial associations between the specifiers, i.e. co-existing but non-ASD specific problems, and functional impairment rather than between symptom severity and functional impairment. This does not provide convincing support for the DSM-5 conceptualization of impairment as determined by symptom severity and that ASD symptoms should be directly linked to an individual's need and level of support.

Globally, our findings add important information to the body of research showing the assets of using standardized diagnostic instruments with good psychometric properties when assessing ASD. We demonstrated levels of objectivity from psychometrically sound to excellent, indeed "almost miraculous" (Kraemer, Kupfer, Clarke, Narrow, & Regier, 2012), in assessing symptoms, criteria and classification with the use of these instruments even in clinical settings. Our results indicate that standardized diagnostic instruments like the ADI-R, perhaps more than any other diagnostic instrument, and the ADOS may contribute to a higher degree of consensus in regards to the assessment of diagnostic criteria of ASD across time, settings and countries. We also showed, like it has been done in previous research reviewed in this thesis, the potential and value of using quantified results of standardized diagnostic instruments to gain new insights, e.g. the association between symptom severity and functional impairment and the DSM-5 ASD specifiers.

Furthermore, our findings together with previous research also indicate that the most serious challenge concerning the use of standardized diagnostic instruments in assessing ASD seems to lie in the overreliance of their results. We showed that, especially for the ADOS in terms of objectivity of classification and the ADI-R in terms of diagnostic validity, the research derived estimates were not directly generalizable to clinical practice. Uncritical use of research derived cutoffs in clinical practice risk to do harm in form misclassification. From our review of previous research, we concluded that we cannot expect to reach complete objectivity nor 100% concurrent sensitivity and specificity with any existing standardized diagnostic instrument, meaning that their results cannot be treated as unambiguous markers of classification. We also concluded that standardized diagnostic instruments, even though potentially more objective in assessing diagnostic criteria than the criteria themselves, suffer from the same validity issues as the diagnoses of the classifications systems they are operationalizing.

6.2 FUTURE DIRECTIONS

Future research of standardized diagnostic instruments should focus on studies on both the objectivity and diagnostic validity in large, heterogeneous but well-described samples of subjects and raters where several aspects are studied in the same sample to further elucidate the instrument's psychometric properties. In terms of optimizing objectivity and comparability of results in research and clinical practice, it would be of great value to continue using the most well established gold standard instruments ADI-R and ADOS. Even with new DSM-5 criteria, the ADI-R and the ADOS have proved their capacity to assess relevant constructs objectively, and they seem to have the potential to continue to do this in regards to future classifications or research paradigms like the RDoC as well.

Future research should continue focusing on how impairment and functioning in ASD and in general should be defined, operationalized and measured, especially in relation to early development, and when diagnosing ASD in infants, toddlers and young preschoolers. Specifically, the component environmental factors, an aspect not explicit in the DSM-5 conceptualization of impairment and not covered in the adaptive functioning construct, need further exploration.

It seems equally important to continue to focus research on how ASD (symptom) severity should be defined, operationalized and measured as ASD symptomatology varies depending on age, IQ and language level and none of the current methods seem to capture this dimensions completely.

7 ACKNOWLEDGEMENTS

8 REFERENCES

American Psychiatric Association [APA]. (1980). *Diagnostic and statistical manual of mental disorders : DSM-III* (Third edition). Washington, DC: American Psychiatric Association.

American Psychiatric Association [APA]. (1994). *Diagnostic and statistical manual of mental disorders : DSM-IV* (Fourth edition). Washington, DC: American Psychiatric Association.

American Psychiatric Association [APA]. (2000). *Diagnostic and statistical manual of mental disorders : DSM-IV-TR* (Fourth edition, text revision). Washington, DC: American Psychiatric Association.

American Psychiatric Association [APA]. (2013). *Diagnostic and statistical manual of mental disorders : DSM-5* (Fifth edition). Arlington, VA: American Psychiatric Association.

Ashwood, K. L., Buitelaar, J., Murphy, D., Spooren, W., & Charman, T. (2015). European clinical network: autism spectrum disorder assessments and patient characterisation. *European Child & Adolescent Psychiatry*, 24(8), 985-995. doi:10.1007/s00787-014-0648-2.

Bailey, A., Le Couteur, A., Gottesman, I., Bolton, P., Simonoff, E., Yuzda, E., & Rutter, M. (1995). Autism as a strongly genetic disorder: evidence from a British twin study. *Psychological Medicine*, 25(1), 63-77. doi:10.1017/S0033291700028099.

Barneveld, P. S., Swaab, H., Fagel, S., van Engeland, H., & de Sonneville, L. M. (2014). Quality of life: a case-controlled long-term follow-up study, comparing young high-functioning adults with autism spectrum disorders with adults with other psychiatric disorders diagnosed in childhood. *Comprehensive Psychiatry*, 55(2), 302-310. doi:10.1016/j.comppsych.2013.08.001.

Barton, M. L., Robins, D. L., Jashar, D., Brennan, L., & Fein, D. (2013). Sensitivity and Specificity of Proposed DSM-5 Criteria for Autism Spectrum Disorder in Toddlers. *Journal of Autism and Developmental Disorders*, 43(5), 1184-1195. doi:10.1007/s10803-013-1817-8.

Bastiaansen, J. A., Meffert, H., Hein, S., Huizinga, P., Ketelaars, C., Pijnenborg, M., Bartels, A., Minderaa, R., Keysers, C., & de Bildt, A. (2011). Diagnosing autism spectrum disorders in adults: the use of Autism Diagnostic Observation Schedule (ADOS) module 4. *Journal of Autism and Developmental Disorders*, 41(9), 1256-1266. doi:10.1007/s10803-010-1157-x.

Bauminger, N., Shulman, C., & Agam, G. (2003). Peer interaction and loneliness in high-functioning children with autism. *Journal of Autism and Developmental Disorders*, 33(5), 489-507. doi:10.1023/A:1025827427901.

Bernier, R. (2012). How do we measure autism severity? SFARI Viewpoints. Retrieved July 2012 from http://sfari.org/news-and-opinion/viewpoint/2012/how-do-we-measure-autism-severity.

Billstedt, E., & Gillberg, C. (2005). Autism after adolescence: population-based 13- to 22-year follow-up study of 120 individuals with autism diagnosed in childhood. *Journal of Autism and Developmental Disorders*, 35(3), 351-360. doi:10.1007/s10803-005-3302-5.

Blacker, D., & Endicott, J. (2000). Psychometric properties: concepts of reliability and validity. In A. J. Rush (Ed.), *Handbook of psychiatric measures* (pp. 7-14). Washington DC: American Psychiatric Association.

Bolduc, M. E., Du Plessis, A. J., Sullivan, N., Khwaja, O. S., Zhang, X., Barnes, K., Robertson, R. L., & Limperopoulos, C. (2011). Spectrum of neurodevelopmental disabilities in children with cerebellar malformations. *Developmental medicine and child neurology*, 53(5), 409-416. doi:10.1111/j.1469-8749.2011.03929.x.

Boucher, J. (2012). Putting theory of mind in its place: psychological explanations of the socio-emotional-communicative impairments in autistic spectrum disorder. *Autism*, 16(3), 226-246. doi:10.1177/1362361311430403.

Bryson, S. E., Zwaigenbaum, L., Brian, J., Roberts, W., Szatmari, P., Rombough, V., & McDermott, C. (2007). A prospective case series of high-risk infants who developed autism. *Journal of Autism and Developmental Disorders*, 37(1), 12-24. doi:10.1007/s10803-006-0328-2.

Buescher, A. V., Cidav, Z., Knapp, M., & Mandell, D. S. (2014). Costs of autism spectrum disorders in the United Kingdom and the United States. *JAMA Pediatrics*, 168(8), 721-728. doi:10.1001/jamapediatrics.2014.210.

Bölte, S. (2012). Brief Report: the Social Responsiveness Scale for Adults (SRS-A): initial results in a German cohort. *Journal of Autism and Developmental Disorders*, 42(9), 1998-1999. doi:10.1007/s10803-011-1424-5.

Bölte, S. (2014). Is autism curable? *Developmental medicine and child neurology*, 56(10), 927-931. doi:10.1111/dmcn.12495.

Bölte, S., de Schipper, E., Robison, J. E., Wong, V. C., Selb, M., Singhal, N., de Vries, P. J., & Zwaigenbaum, L. (2013). Classification of Functioning and Impairment: The Development of ICF Core Sets for Autism Spectrum Disorder. *Autism Research*, 7(1), 167-172. doi:10.1002/aur.1335.

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]. *Zeitschrift für Kinder- und Jugendpsychiatrie und Psychotherapie*, 29(3), 221-229. doi:10.1024//1422-4917.29.3.221.

Bölte, S., & Poustka, F. (2002). The relation between general cognitive level and adaptive behavior domains in individuals with autism with and without co-morbid mental retardation. *Child Psychiatry and Human Development*, 33(2), 165-172. doi:10.1023/A:1020734325815.

Bölte, S., & Poustka, F. (2004). [Diagnostic Observation Scale for Autistic Disorders: initial results of reliability and validity]. *Zeitschrift für Kinder- und Jugendpsychiatrie und Psychotherapie*, 32(1), 45-50. doi:10.1024/1422-4917.32.1.45.

Bölte, S., Poustka, F., & Constantino, J. N. (2008). Assessing autistic traits: cross-cultural validation of the social responsiveness scale (SRS). *Autism Research*, 1(6), 354-363. doi:10.1002/aur.49.

Canino, G. J. (2013). Assessing Child Impairment in Functioning in Different Contexts: Implications for Use of Services and the Classification of Psychiatric Disorders. *Open Journal of Medical Psychology*, 03(01), 29-34. doi:10.4236/ojmp.2013.21006.

Carter, A. S., Garrity-Rokous, F. E., Chazan-Cohen, R., Little, C., & Briggs-Gowan, M. J. (2001). Maternal depression and comorbidity: predicting early parenting, attachment security, and toddler social-emotional problems and competencies. *Journal of the American Academy of Child and Adolescent Psychiatry*, 40(1), 18-26. doi:10.1097/00004583-200101000-00012.

Casanova, M. F. (2014). The Neuropathology of Autism. In F. Volkmar, R. Paul, S. Rogers (Eds.), In F. Volkmar, R. Paul, S. Rogers, & K. Pelphrey (Eds.), *Handbook of Autism and Pervasive Developmental Disorders: Diagnosis, Development, and Brain Mechanisms* (4 ed., Vol. 1, pp. 498-531). Hoboken: Wiley. Retrieved from http://KI.eblib.com/patron/FullRecord.aspx?p=947552.

Centers for Disease Control and Prevention (CDC). (2014). Prevalence of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network, 11 sites, United States, 2010. *MMWR Surveillance Summaries*, 63(2), 1-21. Retrieved from http://www.cdc.gov/mmwr/pdf/ss/ss6302.pdf.

Chakrabarti, S., & Fombonne, E. (2001). Pervasive developmental disorders in preschool children. *JAMA*, 285(24), 3093-3099. doi:10.1001/jama.285.24.3093.

Cicchetti, D. V. (1994). Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology. *Psychological Assessment*, 6(4), 284-290. doi:10.1037/1040-3590.6.4.284.

Cicchetti, D. V., Lord, C., Koenig, K., Klin, A., & Volkmar, F. R. (2008). Reliability of the ADI-R: multiple examiners evaluate a single case. *Journal of Autism and Developmental Disorders*, 38(4), 764-770. doi:10.1007/s10803-007-0448-3.

Cicchetti, D. V., Lord, C., Koenig, K., Klin, A., & Volkmar, F. R. (2014). Reliability of the ADI-R for the single case-part II: clinical versus statistical significance. *Journal of Autism and Developmental Disorders*, 44(12), 3154-3160. doi:10.1007/s10803-014-2177-8.

Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *American journal of mental deficiency*, 86(2), 127-137.

Cicchetti, D. V., Volkmar, F., Klin, A., & Showalter, D. (1995). Diagnosing Autism using ICD-10 criteria: A comparison of neural networks and standard multivariate procedures. *Child Neuropsychology*, 1(1), 26-37. doi:10.1080/09297049508401340.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1), 37-46. doi:10.1177/001316446002000104.

Cohen, J., & Cohen, P. (1983). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (2 ed.). Hillsdale, NJ: Erlbaum.

Colvert, E., Tick, B., McEwen, F., Stewart, C., Curran, S. R., Woodhouse, E., Gillan, N., Hallett, V., Lietz, S., Garnett, T., Ronald, A., Plomin, R., Rijsdijk, F., Happe, F., & Bolton, P. (2015). Heritability of Autism Spectrum Disorder in a UK Population-Based Twin Sample. *JAMA Psychiatry*, 72(5), 415-423. doi:10.1001/jamapsychiatry.2014.3028.

Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, 88(2), 322-328. doi:10.1037/bul0000013.

de Bildt, A., Oosterling, I. J., van Lang, N. D., Kuijper, S., Dekker, V., Sytema, S., Oerlemans, A. M., van Steijn, D. J., Visser, J. C., Rommelse, N. N., Minderaa, R. B., van Engeland, H., van der Gaag, R. J., Buitelaar, J. K., & de Jonge, M. V. (2013). How to Use the ADI-R for Classifying Autism Spectrum Disorders? Psychometric Properties of Criteria from the Literature in 1,204 Dutch Children. *Journal of Autism and Developmental Disorders*, 43(10), 2280-2294. doi:10.1007/s10803-013-1783-1.

de Bildt, A., Sytema, S., Ketelaars, C., Kraijer, D., Mulder, E., Volkmar, F., & Minderaa, R. (2004). Interrelationship between Autism Diagnostic Observation Schedule-Generic (ADOS-G), Autism Diagnostic Interview-Revised (ADI-R), and the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV-TR) classification in children and adolescents with mental retardation. *Journal of Autism and Developmental Disorders*, 34(2), 129-137. doi:10.1023/B:JADD.0000022604.22374.5f.

de Bildt, A., Sytema, S., Meffert, H., & Bastiaansen, J. C. J. (2015). The Autism Diagnostic Observation Schedule, Module 4: Application of the Revised Algorithms in an Independent, Well-Defined, Dutch Sample (n = 93). *Journal of Autism and Developmental Disorders*, 1-10. doi:10.1007/s10803-015-2532-4.

de Bildt, A., Sytema, S., van Lang, N. D., Minderaa, R. B., van Engeland, H., & de Jonge, M. V. (2009). Evaluation of the ADOS revised algorithm: the applicability in 558 Dutch children and adolescents. *Journal of Autism and Developmental Disorders*, 39(9), 1350-1358. doi:10.1007/s10803-009-0749-9.

de Bildt, A., Sytema, S., Zander, E., Bölte, S., Sturm, H., Yirmiya, N., Yaari, M., Charman, T., Salomone, E., LeCouteur, A., Green, J., Bedia, R., Primo, P., van Daalen, E., de Jonge, M., Guðmundsdóttir, E., Jóhannsdóttir, S., Raleva, M., Boskovska, M., Rogé, B., Baduel, S., Moilanen, I., Yliherva, A., Buitelaar, J., & Oosterling, I. (2015). Autism Diagnostic Interview-Revised (ADI-R) Algorithms for Toddlers and Young Preschoolers: Application in a Non-US Sample of 1,104 Children. *Journal of Autism and Developmental Disorders*, 45(7), 1-16 doi:10.1007/s10803-015-2372-2.

de Vet, H. C., Terwee, C. B., Knol, D. L., & Bouter, L. M. (2006). When to use agreement versus reliability measures. *Journal of clinical epidemiology*, 59(10), 1033-1039. doi:10.1016/j.jclinepi.2005.10.015.

DiLavore, P. C., Lord, C., & Rutter, M. (1995). The pre-linguistic autism diagnostic observation schedule. *Journal of Autism and Developmental Disorders*, 25(4), 355-379. doi:10.1007/BF02179373.

Dittmann, V., von Cranach, M., & Eckermann, G. (1990). Abnormalities of adult personality and behaviour (section F 6)--results of the ICD-10 field trial. *Pharmacopsychiatry*, 23 Suppl 4, 170-172. doi:10.1055/s-2007-1014559.

Domellöf, E., Hedlund, L., & Ödman, P. (2014). Health-related quality of life of children and adolescents with functional disabilities in a northern Swedish county. *Quality of life research*, 23(6), 1877-1882. doi:10.1007/s11136-013-0613-4.

Elsabbagh, M., Divan, G., Koh, Y. J., Kim, Y. S., Kauchali, S., Marcin, C., Montiel-Nava, C., Patel, V., Paula, C. S., Wang, C., Yasamy, M. T., & Fombonne, E. (2012). Global prevalence of autism and other pervasive developmental disorders. *Autism Research*, 5(3), 160-179. doi:10.1002/aur.239.

Esler, A., Bal, V., Guthrie, W., Wetherby, A., Weismer, S., & Lord, C. (2015). The Autism Diagnostic Observation Schedule, Toddler Module: Standardized Severity Scores. *Journal of Autism and Developmental Disorders*, 45(9), 2704-2720. doi:10.1007/s10803-015-2432-7.

Fabiano, G. A., & Pelham Jr., W. E. (2009). Impairment in Children. In S. Goldstein & J. A. Naglieri (Eds.), *Assessing Impairment: From Theory to Practice*. Dordrecht: Springer. Retrieved from http://KI.eblib.com/patron/FullRecord.aspx?p=993356.

Fein, D., Barton, M., Eigsti, I. M., Kelley, E., Naigles, L., Schultz, R. T., Stevens, M., Helt, M., Orinstein, A., Rosenthal, M., Troyb, E., & Tyson, K. (2013). Optimal outcome in individuals with a history of autism. *Journal of Child Psychology and Psychiatry*, 54(2), 195-205. doi:10.1111/jcpp.12037.

Fenton, G., D'Ardia, C., Valente, D., Del Vecchio, I., Fabrizi, A., & Bernabei, P. (2003). Vineland adaptive behavior profiles in children with autism and moderate to severe developmental delay. *Autism*, 7(3), 269-287. doi:10.1177/1362361303007003004.

First, M. B., Spitzer, R. L., Gibbon, M., & Williams, J. B. W. (1997). *Structured Clinical Interview for DSM-IV Axis I Disorders SCID-I: Clinician Version, Administration Booklet*: American Psychiatric Press.

Flaum, M., & Andreasen, N. (1995). The reliability of distinguishing primary versus secondary negative symptoms. *Comprehensive Psychiatry*, 36(6), 421-427. doi:10.1016/S0010-440X(95)90249-X.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378-382. doi:10.1037/h0031619.

Freedman, R., Lewis, D. A., Michels, R., Pine, D. S., Schultz, S. K., Tamminga, C. A., Gabbard, G. O., Gau, S. S., Javitt, D. C., Oquendo, M. A., Shrout, P. E., Vieta, E., & Yager, J. (2013). The initial field trials of DSM-5: new blooms and old thorns. *The American journal of psychiatry*, 170(1), 1-5. doi:10.1176/appi.ajp.2012.12091189.

Fulton, E., Eapen, V., Crncec, R., Walter, A., & Rogers, S. (2014). Reducing maladaptive behaviors in preschool-aged children with autism spectrum disorder using the early start denver model. *Frontiers in pediatrics*, 2, 40. doi:10.3389/fped.2014.00040.

Gan, S. M., Tung, L. C., Yeh, C. Y., & Wang, C. H. (2013). ICF-CY based assessment tool for children with autism. *Disability and Rehabilitation*, 35(8), 678-685. doi:10.3109/09638288.2012.705946.

Gathje, R. A., Lewandowski, L. J., & Gordon, M. (2008). The role of impairment in the diagnosis of ADHD. *Journal of Attention Disorders*, 11(5), 529-537. doi:10.1177/1087054707314028.

Geschwind, D. H., & State, M. W. (2015). Gene hunting in autism spectrum disorder: on the path to precision medicine. *Lancet Neurology*, 14(11), 1109-20. doi:10.1016/s1474-4422(15)00044-7.

Gilchrist, A., Green, J., Cox, A., Burton, D., Rutter, M., & Le Couteur, A. (2001). Development and current functioning in adolescents with Asperger syndrome: a comparative study. *Journal of Child Psychology and Psychiatry*, 42(2), 227-240. doi:10.1111/1469-7610.00714.

Gillberg, C., Gillberg, I. C., Råstam, M., & Wentz, E. (2001). The Asperger Syndrome (and high-functioning autism) Diagnostic Interview (ASDI): a preliminary study of a new structured clinical interview. *Autism*, 5(1), 57-66. doi:10.1177/1362361301005001006.

Gillberg, I. C., & Gillberg, C. (1989). Asperger syndrome--some epidemiological considerations: a research note. *Journal of Child Psychology and Psychiatry*, 30(4), 631-638. doi:10.1111/j.1469-7610.1989.tb00275.x.

Goldstein, S., & Naglieri, J. A. (2009). Defining the Evolving Concept of Impairment. In S. Goldstein & J. A. Naglieri (Eds.), *Assessing impairment: From Theory to Practice*. Dordrecht: Springer. Retrieved from http://KI.eblib.com/patron/FullRecord.aspx?p=993356.

Goodman, R., Ford, T., Richards, H., Gatward, R., & Meltzer, H. (2000). The Development and Well-Being Assessment: Description and Initial Validation of an Integrated Assessment of Child and Adolescent Psychopathology. *Journal of Child Psychology and Psychiatry*, 41(5), 645-655. doi:10.1111/j.1469-7610.2000.tb02345.x.

Gotham, K., Pickles, A., & Lord, C. (2009). Standardizing ADOS scores for a measure of severity in autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 39(5), 693-705. doi:10.1007/s10803-008-0674-3.

Gotham, K., Risi, S., Dawson, G., Tager-Flusberg, H., Joseph, R., Carter, A., Hepburn, S., McMahon, W., Rodier, P., Hyman, S. L., Sigman, M., Rogers, S., Landa, R., Spence, M. A., Osann, K., Flodman, P., Volkmar, F., Hollander, E., Buxbaum, J., Pickles, A., & Lord, C. (2008). A replication of the Autism Diagnostic Observation Schedule (ADOS) revised algorithms. *Journal of the American Academy of Child and Adolescent Psychiatry*, 47(6), 642-651. doi:10.1097/CHI.0b013e31816bffb7.

Gotham, K., Risi, S., Pickles, A., & Lord, C. (2007). The Autism Diagnostic Observation Schedule: revised algorithms for improved diagnostic validity. *Journal of Autism and Developmental Disorders*, 37(4), 613-627. doi:10.1007/s10803-006-0280-1.

Gray, K. M., Tonge, B. J., & Sweeney, D. J. (2008). Using the Autism Diagnostic Interview-Revised and the Autism Diagnostic Observation Schedule with young children with developmental delay: evaluating diagnostic validity. *Journal of Autism and Developmental Disorders*, 38(4), 657-667. doi:10.1007/s10803-007-0432-y.

Gullett, J., Donnelly, J. P., Sinert, R., Hosek, B., Fuller, D., Hill, H., Feldman, I., Galetto, G., Auster, M., & Hoffmann, B. (2015). Interobserver agreement in the evaluation of B-lines using bedside ultrasound. *Journal of critical care*, 30(6), 1395-9. doi:10.1016/j.jcrc.2015.08.021.

Gustavsson, A., Svensson, M., Jacobi, F., Allgulander, C., Alonso, J., Beghi, E., Dodel, R., Ekman, M., Faravelli, C., Fratiglioni, L., Gannon, B., Jones, D. H., Jennum, P., Jordanova, A., Jonsson, L., Karampampa, K., Knapp, M., Kobelt, G., Kurth, T., Lieb, R., Linde, M., Ljungcrantz, C., Maercker, A., Melin, B., Moscarelli, M., Musayev, A., Norwood, F., Preisig, M., Pugliatti, M., Rehm, J., Salvador-Carulla, L., Schlehofer, B., Simon, R., Steinhausen, H. C., Stovner, L. J., Vallat, J. M., Van den Bergh, P., van Os, J., Vos, P., Xu, W., Wittchen, H. U., Jonsson, B., & Olesen, J. (2011). Cost of disorders of the brain in Europe 2010. *European neuropsychopharmacology*, 21(10), 718-779. doi:10.1016/j.euroneuro.2011.08.008.

Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in quantitative methods for psychology*, 8(1), 23-34. Retrieved from http://www.tqmp.org/Content/vol08-1/p023/p023.pdf.

Hallmayer, J., Cleveland, S., Torres, A., Phillips, J., Cohen, B., Torigoe, T., Miller, J., Fedele, A., Collins, J., Smith, K., Lotspeich, L., Croen, L. A., Ozonoff, S., Lajonchere, C., Grether, J. K., & Risch, N. (2011). Genetic heritability and shared environmental factors among twin pairs with autism. *Archives of general psychiatry*, 68(11), 1095-1102. doi:10.1001/archgenpsychiatry.2011.76.

Hansen, S. N., Schendel, D. E., & Parner, E. T. (2015). Explaining the increase in the prevalence of autism spectrum disorders: the proportion attributable to changes in reporting practices. *JAMA Pediatrics*, 169(1), 56-62. doi:10.1001/jamapediatrics.2014.1893.

Happé, F., & Frith, U. (2006). The weak coherence account: detail-focused cognitive style in autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 36(1), 5-25. doi:10.1007/s10803-005-0039-0.

Harris, S. L., Handleman, J. S., Belchic, J., & Glasberg, B. (1995). The Vineland Adaptive Behavior Scales for Young Children with Autism. *Special Services in the Schools*, 10(1), 45-54. doi:10.1300/J008v10n01_03.

Harrison, P., & Oakland, T. (2003). *Adaptive Behavior Assessment System™, Second Edition (ABAS-II)*. San Antonio, TX: Harcourt Assessment.

Havdahl, K. A., von Tetzchner, S., Huerta, M., Lord, C., & Bishop, S. L. (2015). Utility of the Child Behavior Checklist as a Screener for Autism Spectrum Disorder. *Autism Research*, 2015 Jul 7 [Epub ahead of print]. doi:10.1002/aur.1515.

Hayes, S. A., & Watson, S. L. (2013). The impact of parenting stress: a meta-analysis of studies comparing the experience of parenting stress in parents of children with and without autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 43(3), 629-642. doi:10.1007/s10803-012-1604-y.

Henninger, N. A., & Taylor, J. L. (2013). Outcomes in adults with autism spectrum disorders: a historical perspective. *Autism*, 17(1), 103-116. doi:10.1177/1362361312441266.

Hill, A., Bölte, S., Petrova, G., Beltcheva, D., Tacheva, S., & Poustka, F. (2001). Stability and interpersonal agreement of the interview-based diagnosis of autism. *Psychopathology*, 34(4), 187-191. doi:10.1159/000049305.

Hirvikoski, T., Jonsson, U., Halldner, L., Lundequist, A., de Schipper, E., Nordin, V., & Bölte, S. (2015). A Systematic Review of Social Communication and Interaction Interventions for Patients with Autism Spectrum Disorder. *Scandinavian Journal of Child and Adolescent Psychiatry and Psychology*, 3(3), 147-168. Retrieved from http://www.sjcapp.org/article/view/22190/19561.

Hofmann, S. G. (2014). Toward a cognitive-behavioral classification system for mental disorders. *Behavior Therapy*, 45(4), 576-587. doi:10.1016/j.beth.2014.03.001.

Howlin, P., Moss, P., Savage, S., & Rutter, M. (2013). Social outcomes in mid- to later adulthood among individuals diagnosed with autism and average nonverbal IQ as children. *Journal of the American Academy of Child and Adolescent Psychiatry*, 52(6), 572-581 e571. doi:10.1016/j.jaac.2013.02.017.

Hsiao, E. Y. (2013). Immune dysregulation in autism spectrum disorder. *International review of neurobiology*, 113, 269-302. doi:10.1016/b978-0-12-418700-9.00009-5.

Huerta, M., Bishop, S. L., Duncan, A., Hus, V., & Lord, C. (2012). Application of DSM-5 Criteria for Autism Spectrum Disorder to Three Samples of Children With DSM-IV Diagnoses of Pervasive Developmental Disorders. *The American journal of psychiatry*, 169(10), 1056-1064. doi:10.1176/appi.ajp.2012.12020276.

Humphrey, N., & Symes, W. (2011). Peer interaction patterns among adolescents with autistic spectrum disorders (ASDs) in mainstream school settings. *Autism*, 15(4), 397-419. doi:10.1177/1362361310387804.

Hus, V., Bishop, S., Gotham, K., Huerta, M., & Lord, C. (2013). Factors influencing scores on the social responsiveness scale. *Journal of Child Psychology and Psychiatry*, 54(2), 216-224. doi:10.1111/j.1469-7610.2012.02589.x.

Hus, V., Gotham, K., & Lord, C. (2014). Standardizing ADOS domain scores: separating severity of social affect and restricted and repetitive behaviors. *Journal of Autism and Developmental Disorders*, 44(10), 2400-2412. doi:10.1007/s10803-012-1719-1.

Hus, V., & Lord, C. (2013). Effects of child characteristics on the Autism Diagnostic Interview-Revised: implications for use of scores as a measure of ASD severity. *Journal of Autism and Developmental Disorders*, 43(2), 371-381. doi:10.1007/s10803-012-1576-y.

Hus, V., & Lord, C. (2014). The Autism Diagnostic Observation Schedule, Module 4: Revised Algorithm and Standardized Severity Scores. *Journal of Autism and Developmental Disorders*, 44(8), 1996-2012. doi:10.1007/s10803-014-2080-3.

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. doi:10.1016/j.biopsych.2006.08.044.

Hyman, S. E. (2010). The diagnosis of mental disorders: the problem of reification. *Annual Review of Clinical Psychology*, 6, 155-179. doi:10.1146/annurev.clinpsy.3.022806.091532.

Idring, S., Lundberg, M., Sturm, H., Dalman, C., Gumpert, C., Rai, D., Lee, B. K., & Magnusson, C. (2015). Changes in prevalence of autism spectrum disorders in 2001-2011: findings from the stockholm youth cohort. *Journal of Autism and Developmental Disorders*, 45(6), 1766-1773. doi:10.1007/s10803-014-2336-y.

Idring, S., Rai, D., Dal, H., Dalman, C., Sturm, H., Zander, E., Lee, B. K., Serlachius, E., & Magnusson, C. (2012). Autism spectrum disorders in the stockholm youth cohort: design, prevalence and validity. *PLoS One*, 7(7), e41280. doi:10.1371/journal.pone.0041280.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., Sanislow, C., & Wang, P. (2010). Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *The American journal of psychiatry*, 167(7), 748-751. doi:10.1176/appi.ajp.2010.09091379.

Jacob, A., Scott, M., Falkmer, M., & Falkmer, T. (2015). The Costs and Benefits of Employing an Adult with Autism Spectrum Disorder: A Systematic Review. *PLoS One*, 10(10), e0139896. doi:10.1371/journal.pone.0139896.

Joshi, G., Petty, C. R., Fried, R., Wozniak, J., Micco, J. A., Henin, A., Doyle, R., Galdo, M., Kotarski, M., Caruso, J., Meller, B., Faraone, S. V., & Biederman, J. (2011). Discriminant and concurrent validity of a simplified DSM-based structured diagnostic instrument for the assessment of autism spectrum disorders in youth and young adults. *BMC Psychiatry*, 11, 204. doi:10.1186/1471-244x-11-204.

Järbrink, K. (2007). The economic consequences of autistic spectrum disorder among children in a Swedish municipality. *Autism*, 11(5), 453-463. doi:10.1177/1362361307079602.

Kanne, S. M., Gerber, A. J., Quirmbach, L. M., Sparrow, S. S., Cicchetti, D. V., & Saulnier, C. A. (2011). The role of adaptive behavior in autism spectrum disorders: implications for functional outcome. *Journal of Autism and Developmental Disorders*, 41(8), 1007-1018. doi:10.1007/s10803-010-1126-4.

Kanner, L. (1943). Autistic disturbances of affective contact. *Nervous Child*, 2, 217-250. Retrieved from http://www.neurodiversity.com/library_kanner_1943.pdf.

Kaufman, J., Birmaher, B., Brent, B., Rao, U. M. A., Ryan, N., Axelson, D., Zelazny, J., & Gill, M. K. (2009). Kiddie-SADS-Present and Lifetime Version (K-SADS-PL), 2009 Working Draft. Retrieved July 23, 2015 from

http://www.psychiatry.pitt.edu/sites/default/files/Documents/assessments/KSADS-PL_2009_working_draft_full.pdf.

Kaufman, J., Birmaher, B., Brent, D., Rao, U., Flynn, C., Moreci, P., Williamson, D., & Ryan, N. (1997). Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime Version (K-SADS-PL): Initial Reliability and Validity Data. *Journal of the American Academy of Child and Adolescent Psychiatry*, 36(7), 980-988. doi:10.1097/00004583-199707000-00021.

Kim, S. H., & Lord, C. (2012a). Combining information from multiple sources for the diagnosis of autism spectrum disorders for toddlers and young preschoolers from 12 to 47 months of age. *Journal of Child Psychology and Psychiatry*, 53(2), 143-151. doi:10.1111/j.1469-7610.2011.02458.x.

Kim, S. H., & Lord, C. (2012b). New autism diagnostic interview-revised algorithms for toddlers and young preschoolers from 12 to 47 months of age. *Journal of Autism and Developmental Disorders*, 42(1), 82-93. doi:10.1007/s10803-011-1213-1.

Kim, S. H., Thurm, A., Shumway, S., & Lord, C. (2013). Multisite Study of New Autism Diagnostic Interview-Revised (ADI-R) Algorithms for Toddlers and Young Preschoolers. *Journal of Autism and Developmental Disorders*, 43(7), 1527-1538. doi:10.1007/s10803-012-1696-4.

Kim, Y. S., Leventhal, B. L., Koh, Y. J., Fombonne, E., Laska, E., Lim, E. C., Cheon, K. A., Kim, S. J., Kim, Y. K., Lee, H., Song, D. H., & Grinker, R. R. (2011). Prevalence of autism spectrum disorders in a total population sample. *The American journal of psychiatry*, 168(9), 904-912. doi:10.1176/appi.ajp.2011.10101532.

Klein-Tasman, B. P., Risi, S., & Lord, C. E. (2007). Effect of language and task demands on the diagnostic effectiveness of the autism diagnostic observation schedule: the impact of module choice. *Journal of Autism and Developmental Disorders*, 37(7), 1224-1234. doi:10.1007/s10803-006-0266-z.

Klin, A., Lang, J., Cicchetti, D. V., & Volkmar, F. R. (2000). Brief report: Interrater reliability of clinical diagnosis and DSM-IV criteria for autistic disorder: results of the DSM-IV autism field trial. *Journal of Autism and Developmental Disorders*, 30(2), 163-167. doi:10.1023/A:1005415823867.

Klin, A., Saulnier, C., Chawarska, K., & Volkmar, F. R. (2008). Case studies of infants first evaluated in the second year of life. In K. Chawarska, A. Klin & F. Volkmar (Eds.), Autism spectrum disorders in infants and toddlers: Diagnosis, assessment, and treatment (pp. 141–169). New York: Guilford.

Klin, A., Saulnier, C. A., Sparrow, S. S., Cicchetti, D. V., Volkmar, F. R., & Lord, C. (2007). Social and communication abilities and disabilities in higher functioning individuals with autism spectrum disorders: the Vineland and the ADOS. *Journal of Autism and Developmental Disorders*, 37(4), 748-759. doi:10.1007/s10803-006-0229-4.

Knapp, M., Romeo, R., & Beecham, J. (2009). Economic cost of autism in the UK. *Autism*, 13(3), 317-336. doi:10.1177/1362361309104246.

Koran, L. M. (1975a). The reliability of clinical methods, data and judgments (first of two parts). *The New England journal of medicine*, 293(13), 642-646. doi:10.1056/NEJM197509252931307.

Koran, L. M. (1975b). The reliability of clinical methods, data and judgments (second of two parts). *The New England journal of medicine*, 293(14), 695-701. doi:10.1056/NEJM197510022931405.

Kraemer, H. C., Kupfer, D. J., Clarke, D. E., Narrow, W. E., & Regier, D. A. (2012). DSM-5: how reliable is reliable enough? *The American journal of psychiatry*, 169(1), 13-15. doi:10.1176/appi.ajp.2011.11010050.

Kraijer, D. (2000). Review of adaptive behavior studies in mentally retarded persons with autism/pervasive developmental disorder. *Journal of Autism and Developmental Disorders*, 30(1), 39-47. doi:10.1023/A:1005460027636.

Kuhlthau, K., Orlich, F., Hall, T. A., Sikora, D., Kovacs, E. A., Delahaye, J., & Clemons, T. E. (2010). Health-Related Quality of Life in children with autism spectrum disorders: results from the autism treatment network. *Journal of Autism and Developmental Disorders*, 40(6), 721-729. doi:10.1007/s10803-009-0921-2.

Kulage, K. M., Smaldone, A. M., & Cohn, E. G. (2014). How Will DSM-5 Affect Autism Diagnosis? A Systematic Literature Review and Meta-analysis. *Journal of Autism and Developmental Disorders*, 44(8), 1918-1932. doi:10.1007/s10803-014-2065-2.

Kushnirsky, M., Nguyen, V., Katz, J. S., Steinklein, J., Rosen, L., Warshall, C., Schulder, M., & Knisely, J. P. (2015). Time-delayed contrast-enhanced MRI improves detection of brain metastases and apparent treatment volumes. *Journal of neurosurgery*, 1-7. doi:10.3171/2015.2.jns141993.

Lai, M. C., Lombardo, M. V., & Baron-Cohen, S. (2013). Autism. *Lancet*, 383(9920), 896-910. doi:10.1016/s0140-6736(13)61539-1.

Lam, K. S., Aman, M. G., & Arnold, L. E. (2006). Neurochemical correlates of autistic disorder: a review of the literature. *Research in developmental disabilities*, 27(3), 254-289. doi:10.1016/j.ridd.2005.03.003.

Lampi, K. M., Sourander, A., Gissler, M., Niemela, S., Rehnstrom, K., Pulkkinen, E., Peltonen, L., & Von Wendt, L. (2010). Brief report: validity of Finnish registry-based diagnoses of autism with the ADI-R. *Acta paediatrica*, 99(9), 1425-1428. doi:10.1111/j.1651-2227.2010.01835.x.

Lasgaard, M., Nielsen, A., Eriksen, M. E., & Goossens, L. (2010). Loneliness and social support in adolescent boys with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 40(2), 218-226. doi:10.1007/s10803-009-0851-z.

Lavelle, T. A., Weinstein, M. C., Newhouse, J. P., Munir, K., Kuhlthau, K. A., & Prosser, L. A. (2014). Economic burden of childhood autism spectrum disorders. *Pediatrics*, 133(3), e520-529. doi:10.1542/peds.2013-0763.

Le Couteur, A., Haden, G., Hammal, D., & McConachie, H. (2008). Diagnosing autism spectrum disorders in pre-school children using two standardised assessment instruments: the ADI-R and the ADOS. *Journal of Autism and Developmental Disorders*, 38(2), 362-372. doi:10.1007/s10803-007-0403-3.

Le Couteur, A., Rutter, M., Lord, C., Rios, P., Robertson, S., Holdgrafer, M., & McLennan, J. (1989). Autism diagnostic interview: a standardized investigator-based instrument. *Journal of Autism and Developmental Disorders*, 19(3), 363-387. doi:10.1007/BF02212936.

Lecavalier, L., Aman, M. G., Scahill, L., McDougle, C. J., McCracken, J. T., Vitiello, B., Tierney, E., Arnold, L. E., Ghuman, J. K., Loftin, R. L., Cronin, P., Koenig, K., Posey, D. J., Martin, A., Hollway, J., Lee, L. S., & Kau, A. S. (2006). Validity of the autism diagnostic interview-revised. *American Journal on Mental Retardation*, 111(3), 199-215. doi:10.1352/0895-8017(2006)111[199:VOTADI]2.0.CO;2.

Leekam, S. R., Libby, S. J., Wing, L., Gould, J., & Taylor, C. (2002). The Diagnostic Interview for Social and Communication Disorders: algorithms for ICD-10 childhood autism and Wing and Gould autistic spectrum disorder. *Journal of Child Psychology and Psychiatry*, 43(3), 327-342. doi:10.1111/1469-7610.00024.

Leigh, J. P., & Du, J. (2015). Brief Report: Forecasting the Economic Burden of Autism in 2015 and 2025 in the United States. *Journal of Autism and Developmental Disorders,* 2015 Jul 17 [Epub ahead of print]. doi:10.1007/s10803-015-2521-7.

Leung, R. C., & Zakzanis, K. K. (2014). Brief report: cognitive flexibility in autism spectrum disorders: a quantitative review. *Journal of Autism and Developmental Disorders*, 44(10), 2628-2645. doi:10.1007/s10803-014-2136-4.

Lichtenstein, P., Carlström, E., Råstam, M., Gillberg, C., & Anckarsäter, H. (2010). The Genetics of Autism Spectrum Disorders and Related Neuropsychiatric Disorders in Childhood. *The American journal of psychiatry*, 167(11), 1357-1363. doi:10.1176/appi.ajp.2010.10020223.

Lienert, G. (1989). *Testaufbau und Testanalyse* (4 ed.). München - Weinheim: Psychologie Verlags Union.

Liss, M., Harel, B., Fein, D., Allen, D., Dunn, M., Feinstein, C., Morris, R., Waterhouse, L., & Rapin, I. (2001). Predictors and correlates of adaptive functioning in children with developmental disorders. *Journal of Autism and Developmental Disorders*, 31(2), 219-230. doi:10.1023/A:1010707417274.

London, E. B. (2014). Categorical diagnosis: a fatal flaw for autism research? *Trends in neurosciences*, 37(12), 683-686. doi:10.1016/j.tins.2014.10.003.

Lord, C., Corsello, C., & Grzadzinski, R. (2014). Diagnostic Instruments in Autistic Spectrum Disorder. In F. Volkmar, R. Paul & S. Rogers (Eds.), *Handbook of Autism and Pervasive Developmental Disorders: Assessment, Interventions, and Policy* (4 ed., Vol. 2, pp. 609-660). Hoboken: Wiley. Retrieved from http://KI.eblib.com/patron/FullRecord.aspx?p=947857.

Lord, C., Pickles, A., McLennan, J., Rutter, M., Bregman, J., Folstein, S., Fombonne, E., Leboyer, M., & Minshew, N. (1997). Diagnosing autism: analyses of data from the Autism Diagnostic Interview. *Journal of Autism and Developmental Disorders*, 27(5), 501-517. doi:10.1023/A:1025873925661.

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Jr., Leventhal, B. L., 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(3), 205-223. doi:10.1023/A:1005592401947.

Lord, C., Rutter, M., DiLavore, P., & Risi, S. (1999). *ADOS. Autism Diagnostic Observation Schedule*. Los Angeles, CA: Western Psychological Publishing.

Lord, C., Rutter, M., DiLavore, P., Risi, S., Gotham, K., & Bishop, S. (2012). *Autism Diagnostic Observation Schedule, Second Edition (ADOS-2). Manual (Part I)*. Torrance, CA: Western Psychological Services.

Lord, C., Rutter, M., Goode, S., Heemsbergen, J., Jordan, H., Mawhood, L., & Schopler, E. (1989). Autism diagnostic observation schedule: a standardized observation of communicative and social behavior. *Journal of Autism and Developmental Disorders*, 19(2), 185-212. doi:10.1007/BF02211841.

Lord, C., Rutter, M., & Le Couteur, A. (1994). Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders*, 24(5), 659-685. doi:10.1007/BF02172145.

Lord, C., Storoschuk, S., Rutter, M., & Pickles, A. (1993). Using the ADI-R to diagnose autism in preschool children. *Infant Mental Health Journal*, 14(3), 234-252. doi:10.1002/1097-0355(199323)14:3<234::AID-IMHJ2280140308>3.0.CO;2-F.

Lotter, V. (1966). Epidemiology of autistic conditions in young children. *Social Psychiatry*, 1, 124-137. doi:10.1007/BF00584048.

Luyster, R., Gotham, K., Guthrie, W., Coffing, M., Petrak, R., Pierce, K., Bishop, S., Esler, A., Hus, V., Oti, R., Richler, J., Risi, S., & Lord, C. (2009). The Autism Diagnostic Observation Schedule-toddler module: a new module of a standardized diagnostic measure for autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 39(9), 1305-1320. doi:10.1007/s10803-009-0746-z [doi].

Lyall, K., Schmidt, R. J., & Hertz-Picciotto, I. (2014). Environmental Factors in the Preconception and Prenatal Periods in Relation to Risk for ASD. In F. Volkmar, R. Paul, S. Rogers, & K. Pelphrey (Eds.), *Handbook of Autism and Pervasive Developmental Disorders: Diagnosis, Development, and Brain Mechanisms* (4 ed., Vol. 1, pp. 424-456). Hoboken: Wiley. Retrieved from http://KI.eblib.com/patron/FullRecord.aspx?p=947552.

Maenner, M. J., Rice, C. E., Arneson, C. L., Cunniff, C., Schieve, L. A., Carpenter, L. A., Van Naarden Braun, K., Kirby, R. S., Bakian, A. V., & Durkin, M. S. (2014). Potential impact of DSM-5 criteria on autism spectrum disorder prevalence estimates. *JAMA Psychiatry*, 71(3), 292-300. doi:10.1001/jamapsychiatry.2013.3893.

Magiati, I., Tay, X. W., & Howlin, P. (2014). Cognitive, language, social and behavioural outcomes in adults with autism spectrum disorders: A systematic review of longitudinal follow-up studies in adulthood. *Clinical Psychology Review*, 34(1), 73-86. doi:10.1016/j.cpr.2013.11.002.

Mahoney, W. J., Szatmari, P., MacLean, J. E., Bryson, S. E., Bartolucci, G., Walter, S. D., Jones, M. B., & Zwaigenbaum, L. (1998). Reliability and accuracy of differentiating pervasive developmental disorder subtypes. *Journal of the American Academy of Child and Adolescent Psychiatry*, 37(3), 278-285. doi:10.1097/00004583-199803000-00012.

Majnemer, A., Limperopoulos, C., Shevell, M., Rohlicek, C., Rosenblatt, B., & Tchervenkov, C. (2012). Gender differences in the developmental outcomes of children with congenital cardiac defects. *Cardiology in the Young*, 22(5), 514-519. doi:10.1017/s1047951111002071.

Mazefsky, C. A., & Oswald, D. P. (2006). The discriminative ability and diagnostic utility of the ADOS-G, ADI-R, and GARS for children in a clinical setting. *Autism*, 10(6), 533-549. doi:10.1177/1362361306068505.

McEvilly, M., Wicks, S., & Dalman, C. (2015). Sick Leave and Work Participation Among Parents of Children with Autism Spectrum Disorder in the Stockholm Youth Cohort: A Register Linkage Study in Stockholm, Sweden. *Journal of Autism and Developmental Disorders*, 45(7), 2157-2167. doi:10.1007/s10803-015-2381-1.

Molloy, C. A., Murray, D. S., Akers, R., Mitchell, T., & Manning-Courtney, P. (2011). Use of the Autism Diagnostic Observation Schedule (ADOS) in a clinical setting. *Autism*, 15(2), 143-162. doi:10.1177/1362361310379241.

Mullen, E. M. (1995). *Mullen scales of early learning (AGS ed.)*. Circle Pines, MN: American Guidance Service.

Naglieri, J. A. (2009). Psychometric Issues in the Assessment of Impairment. In S. Goldstein & J. A. Naglieri (Eds.), Assessing Impairment : From Theory to Practice. Dordrecht: Springer. Retrieved from http://KI.eblib.com/patron/FullRecord.aspx?p=993356.

Nah, Y. H., Young, R. L., Brewer, N., & Berlingeri, G. (2014). Autism Detection in Early Childhood (ADEC): reliability and validity data for a Level 2 screening tool for autistic disorder. *Psychological Assessment*, 26(1), 215-226. doi:10.1037/a0034472.

Ness, S., Todd, M. J., Wang, S., Eerdekens, M., Nye, J. S., & Ford, L. (2012). Adaptive behavior outcomes in infants treated with adjunctive topiramate. *Pediatric Neurology*, 46(6), 350-358. doi:10.1016/j.pediatrneurol.2012.02.028.

Newcombe, R. G. (1998). Interval estimation for the difference between independent proportions: Comparison of eleven methods. *Statistics in Medicine*, 17, 873-890. doi:10.1002/(SICI)1097-0258(19980430)17:8<873::AID-SIM779>3.0.CO;2-I.

NICE Clinical Guidelines. (2011). *Autism: Recognition, Referral, Diagnosis and Management of Children and Young People on the Autism Spectrum*. Leicester UK: The British Psychological Society & The Royal College of Psychiatrists, 2011. Retrieved from http://www.guideline.gov/content.aspx?id=34822.

Noriega, D. B., & Savelkoul, H. F. (2014). Immune dysregulation in autism spectrum disorder. *European journal of pediatrics*, 173(1), 33-43. doi:10.1007/s00431-013-2183-4.

Nygren, G., Hagberg, B., Billstedt, E., Skoglund, A., Gillberg, C., & Johansson, M. (2009). The Swedish version of the Diagnostic Interview for Social and Communication Disorders (DISCO-10). Psychometric properties. *Journal of Autism and Developmental Disorders*, 39(5), 730-741. doi:10.1007/s10803-008-0678-z.

Olson, D. M., Stutzman, S., Saju, C., Wilson, M., Zhao, W., & Aiyagari, V. (2015). Interrater Reliability of Pupillary Assessments. *Neurocritical Care,* 2015 Sep 17 [Epub ahead of print]. doi:10.1007/s12028-015-0182-1.

Oosterling, I., Roos, S., de Bildt, A., Rommelse, N., de Jonge, M., Visser, J., Lappenschaar, M., Swinkels, S., van der Gaag, R. J., & Buitelaar, J. (2010). Improved diagnostic validity of the ADOS revised algorithms: a replication study in an independent sample. *Journal of Autism and Developmental Disorders*, 40(6), 689-703. doi:10.1007/s10803-009-0915-0.

Ou, J. J., Shi, L. J., Xun, G. L., Chen, C., Wu, R. R., Luo, X. R., Zhang, F. Y., & Zhao, J. P. (2015). Employment and financial burden of families with preschool children diagnosed with autism spectrum disorders in urban China: results from a descriptive study. *BMC Psychiatry*, 15, 3. doi:10.1186/s12888-015-0382-4.

Overbury, R., Murtaugh, M. A., Fischer, A., & Frech, T. M. (2015). Primary care assessment of capillaroscopy abnormalities in patients with Raynaud's phenomenon. *Clinical rheumatology*, 2015 Sep 25 [Epub ahead of print]. doi:10.1007/s10067-015-3062-3.

Ozonoff, S., Young, G. S., Belding, A., Hill, M., Hill, A., Hutman, T., Johnson, S., Miller, M., Rogers, S. J., Schwichtenberg, A. J., Steinfeld, M., & Iosif, A. M. (2014). The broader autism phenotype in infancy: when does it emerge? *Journal of the American Academy of Child and Adolescent Psychiatry*, 53(4), 398-407 e392. doi:10.1016/j.jaac.2013.12.020.

Papanikolaou, K., Paliokosta, E., Houliaras, G., Vgenopoulou, S., Giouroukou, E., Pehlivanidis, A., Tomaras, V., & Tsiantis, I. (2009). Using the Autism Diagnostic Interview-Revised and the Autism Diagnostic Observation Schedule-Generic for the diagnosis of autism spectrum disorders in a Greek sample with a wide range of intellectual abilities. *Journal of Autism and Developmental Disorders*, 39(3), 414-420. doi:10.1007/s10803-008-0639-6.

Papazoglou, A., Jacobson, L. A., & Zabel, T. A. (2013). Sensitivity of the BASC-2 Adaptive Skills Composite in detecting adaptive impairment in a clinically referred sample of children and adolescents. *Clinical Neuropsychologist*, 27(3), 386-395. doi:10.1080/13854046.2012.760651.

Park, C. J., Yelland, G. W., Taffe, J. R., & Gray, K. M. (2012). Brief report: The relationship between language skills, adaptive behavior, and emotional and behavior problems in pre-schoolers with autism. *Journal of Autism and Developmental Disorders*, 42(12), 2761-2766. doi:10.1007/s10803-012-1534-8.

Paul, R., Loomis, R., & Chawarska, K. (2011). Adaptive Behavior in Toddlers Under Two with Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders,* 44(2), 264-270. doi:10.1007/s10803-011-1279-9.

Paul, R., Miles, S., Cicchetti, D., Sparrow, S., Klin, A., Volkmar, F., Coflin, M., & Booker, S. (2004). Adaptive behavior in autism and Pervasive Developmental Disorder-Not Otherwise Specified: microanalysis of scores on the Vineland Adaptive Behavior Scales. *Journal of Autism and Developmental Disorders*, 34(2), 223-228. doi:10.1023/B:JADD.0000022612.18116.46.

Pelphrey, K. A., Shultz, S., Hudac, C. M., & Vander Wyk, B. C. (2011). Research Review: Constraining heterogeneity: the social brain and its development in autism spectrum disorder. *Journal of Child Psychology and Psychiatry*, 52(6), 631-44. doi:10.1111/j.1469-7610.2010.02349.x.

Perry, A., Flanagan, H. E., Dunn Geier, J., & Freeman, N. L. (2009). Brief report: the Vineland Adaptive Behavior Scales in young children with autism spectrum disorders at different cognitive levels. *Journal of Autism and Developmental Disorders*, 39(7), 1066-1078. doi:10.1007/s10803-009-0704-9.

Philip, R. C., Dauvermann, M. R., Whalley, H. C., Baynham, K., Lawrie, S. M., & Stanfield, A. C. (2012). A systematic review and meta-analysis of the fMRI investigation of autism spectrum disorders. *Neuroscience and biobehavioral reviews*, 36(2), 901-942. doi:10.1016/j.neubiorev.2011.10.008.

Poustka, F., Lisch, S., Ruhl, D., Sacher, A., Schmotzer, G., & Werner, K. (1996). The standardized diagnosis of autism, Autism Diagnostic Interview-Revised: interrater reliability of the German form of the interview. *Psychopathology*, 29(3), 145-153. doi:10.1159/000284984.

Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: implications for estimating interrater reliability. *The Journal of applied psychology*, 93(5), 959-981. doi:10.1037/0021-9010.93.5.959.

Quigg, M., Sun, F., Fountain, N. B., Jobst, B. C., Wong, V. S., Mirro, E., Brown, S., & Spencer, D. C. (2015). Interrater reliability in interpretation of electrocorticographic seizure detections of the responsive neurostimulator. *Epilepsia*, 56(6), 968-971. doi:10.1111/epi.12998.

Rane, P., Cochran, D., Hodge, S. M., Haselgrove, C., Kennedy, D. N., & Frazier, J. A. (2015). Connectivity in Autism: A Review of MRI Connectivity Studies. *Harvard review of psychiatry*, 23(4), 223-244. doi:10.1097/hrp.0000000000000072.

Rapee, R. M., Bőgels, S. M., van der Sluis, C. M., Craske, M. G., & Ollendick, T. (2012). Annual Research Review: Conceptualising functional impairment in children and adolescents. *Journal of Child Psychology and Psychiatry*, 53(5), 454-468. doi:10.1111/j.1469-7610.2011.02479.x.

Ray-Subramanian, C. E., Huai, N., & Ellis Weismer, S. (2011). Brief report: adaptive behavior and cognitive skills for toddlers on the autism spectrum. *Journal of Autism and Developmental Disorders*, 41(5), 679-684. doi:10.1007/s10803-010-1083-y.

Regier, D. A., Kaelber, C. T., Roper, M. T., Rae, D. S., & Sartorius, N. (1994). The ICD-10 clinical field trial for mental and behavioral disorders: results in Canada and the United States. *The American journal of psychiatry*, 151(9), 1340-1350. doi:10.1176/ajp.151.9.1340.

Reichow, B. (2012). Overview of meta-analyses on early intensive behavioral intervention for young children with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 42(4), 512-520. doi:10.1007/s10803-011-1218-9.

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. doi:10.1097/01.chi.0000227880.42780.0e.

Roid, G. H., & Sampers, J. L. (2005). *Merrill-Palmer – Revised Scales of Development (Swedish version)*: Stoelting.

Rutter, M. (1978). Diagnosis and definition of childhood autism. *Journal of Autism and Childhood Schizophrenia*, 8(2), 139-161. doi:10.1007/BF01537863.

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

Saemundsen, E., Magnusson, P., Smari, J., & Sigurdardottir, S. (2003). Autism Diagnostic Interview-Revised and the Childhood Autism Rating Scale: convergence and discrepancy in diagnosing autism. *Journal of Autism and Developmental Disorders*, 33(3), 319-328. doi:10.1023/A:1024410702242.

Samson, F., Mottron, L., Soulières, I., & Zeffiro, T. A. (2012). Enhanced visual functioning in autism: an ALE meta-analysis. *Human brain mapping*, 33(7), 1553-1581. doi:10.1002/hbm.21307.

Sandin, S., Lichtenstein, P., Kuja-Halkola, R., Larsson, H., Hultman, C. M., & Reichenberg, A. (2014). The familial risk of autism. *JAMA*, 311(17), 1770-1777. doi:10.1001/jama.2014.4144.

Sappok, T., Diefenbacher, A., Budczies, J., Schade, C., Grubich, C., Bergmann, T., Bölte, S., & Dziobek, I. (2013). Diagnosing autism in a clinical sample of adults with intellectual disabilities: how useful are the ADOS and the ADI-R? *Research in developmental disabilities*, 34(5), 1642-1655. doi:10.1016/j.ridd.2013.01.028.

Sartorius, N., Üstün, T. B., Korten, A., Cooper, J. E., & van Drimmelen, J. (1995). Progress toward achieving a common language in psychiatry, II: Results from the international field trials of the ICD-10 diagnostic criteria for research for mental and behavioral disorders. *The American journal of psychiatry*, 152(10), 1427-1437. doi:10.1176/ajp.152.10.1427.

Saulnier, C. A., & Klin, A. (2007). Brief report: social and communication abilities and disabilities in higher functioning individuals with autism and Asperger syndrome. *Journal of Autism and Developmental Disorders*, 37(4), 788-793. doi:10.1007/s10803-006-0288-6.

Schopler, E., Lansing, M. D., Reichler, R. J., & Marcus, L. M. (2005). *Psychoeducational profile: TEACCH individualized psychoeducational assessment for children with autism spectrum disorders* (3 ed.). Austin, TX: PRO-ED.

Schroeder, J. H., Cappadocia, M. C., Bebko, J. M., Pepler, D. J., & Weiss, J. A. (2014). Shedding light on a pervasive problem: a review of research on bullying

experiences among children with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 44(7), 1520-1534. doi:10.1007/s10803-013-2011-8.

Senju, A. (2012). Spontaneous theory of mind and its absence in autism spectrum disorders. *Neuroscientist*, 18(2), 108-113. doi:10.1177/1073858410397208.

Shaffer, D., Gould, M. S., Brasic, J., Ambrosini, P., Fisher, P., Bird, H., & Aluwahlia, S. (1983). A children's global assessment scale (CGAS). *Archives of general psychiatry*, 40(11), 1228-1231. doi:10.1001/archpsyc.1983.01790100074010.

Shavelson, R. J., & Webb, N. M. (2006). Generalizability Theory. In *Handbook of complementary methods in education research* (pp. 309-322). Mahwah, NJ: Lawrence Erlbaum Associates Publishers; US.

Simonoff, E., Jones, C. R., Baird, G., Pickles, A., Happé, F., & Charman, T. (2013). The persistence and stability of psychiatric problems in adolescents with autism spectrum disorders. *Journal of Child Psychology and Psychiatry*, 54(2), 186-194. doi:10.1111/j.1469-7610.2012.02606.x.

Skuse, D., Warrington, R., Bishop, D., Chowdhury, U., Lau, J., Mandy, W., & Place, M. (2004). The developmental, dimensional and diagnostic interview (3di): a novel computerized assessment for autism spectrum disorders. *Journal of the American Academy of Child and Adolescent Psychiatry*, 43(5), 548-558. doi:10.1097/00004583-200405000-00008.

Smith, I., Reichow, B., & Volkmar, F. (2015). The Effects of DSM-5 Criteria on Number of Individuals Diagnosed with Autism Spectrum Disorder: A Systematic Review. *Journal of Autism and Developmental Disorders*, 45(8), 2541-2552. doi:10.1007/s10803-015-2423-8.

Smith, L. E., Greenberg, J. S., & Mailick, M. R. (2012). Adults with autism: outcomes, family effects, and the multi-family group psychoeducation model. *Current psychiatry reports*, 14(6), 732-738. doi:10.1007/s11920-012-0328-1.

Sparrow, S., Balla, D. A., & Cichetti, D. V. (1984). *Vineland Adaptive Behavior Scales (Vineland ABS)*. Circle Pine, MN: American Guidance Service.

Sparrow, S., Cichetti, D. V., & Balla, D. A. (2005). *Vineland-II. Vineland Adaptive Behavior Scales*. (2 ed.). Circle Pines, MN: AGS Publishing.

Spitzer, R. L. (1983). Psychiatric diagnosis: Are clinicians still necessary? *Comprehensive Psychiatry*, 24(5), 399-411. doi:1016/0010-440X(83)90032-9.

Stone, W. L., Coonrod, E. E., Pozdol, S. L., & Turner, L. M. (2003). The Parent Interview for Autism-Clinical Version (PIA-CV): a measure of behavioral change for young children with autism. *Autism*, 7(1), 9-30. doi:10.1177/1362361303007001003.

Stone, W. L., & Hogan, K. L. (1993). A structured parent interview for identifying young children with autism. *Journal of Autism and Developmental Disorders*, 23(4), 639-652. doi:10.1007/BF01046106.

Sturmey, P., & Dalfern, S. (2014). The Effects of DSM5 Autism Diagnostic Criteria on Number of Individuals Diagnosed with Autism Spectrum Disorders: A Systematic Review. *Review Journal of Autism and Developmental Disorders*, 1(4), 249-252. doi:10.1007/s40489-014-0016-7.

Sutera, S., Pandey, J., Esser, E. L., Rosenthal, M. A., Wilson, L. B., Barton, M., Green, J., Hodgson, S., Robins, D. L., Dumont-Mathieu, T., & Fein, D. (2007). Predictors of optimal outcome in toddlers diagnosed with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 37(1), 98-107. doi:10.1007/s10803-006-0340-6.

Swedish Council on Health Technology Assessment [Statens beredning för medicinsk utvärdering, SBU]. (2013). [*Autism Spectrum Disorder. Diagnostic Assessment and Interventions, Organization of Services and Participation – A Systematic Review*] (Vol. 215). Retrieved from http://www.sbu.se/upload/Publikationer/Content0/1/Autismspektrumtillst%C3%A5nd_fulltext.pdf.

Szatmari, P., Bryson, S. E., Boyle, M. H., Streiner, D. L., & Duku, E. (2003). Predictors of outcome among high functioning children with autism and Asperger syndrome. *Journal of Child Psychology and Psychiatry*, 44(4), 520-528. doi:10.1111/1469-7610.00141.

Tanguay, P. E., Robertson, J., & Derrick, A. (1998). A dimensional classification of autism spectrum disorder by social communication domains. *Journal of the American Academy of Child and Adolescent Psychiatry*, 37(3), 271-277. doi:10.1097/00004583-199803000-00011.

The National Board of Health and Welfare [Socialstyrelsen]. (2009). [*Methods used in Child and Adolescent Psychiatry. A National Inventory*]. Retrieved from http://www.socialstyrelsen.se/Lists/Artikelkatalog/Attachments/8367/2009-126-146_2009126146.pdf.

Tomanik, S. S., Pearson, D. A., Loveland, K. A., Lane, D. M., & Bryant Shaw, J. (2007). Improving the reliability of autism diagnoses: examining the utility of adaptive behavior. *Journal of Autism and Developmental Disorders*, 37(5), 921-928. doi:10.1007/s10803-006-0227-6.

Tsuchiya, K. J., Matsumoto, K., Yagi, A., Inada, N., Kuroda, M., Inokuchi, E., Koyama, T., Kamio, Y., Tsujii, M., Sakai, S., Mohri, I., Taniike, M., Iwanaga, R., Ogasahara, K., Miyachi, T., Nakajima, S., Tani, I., Ohnishi, M., Inoue, M., Nomura, K., Hagiwara, T., Uchiyama, T., Ichikawa, H., Kobayashi, S., Miyamoto, K., Nakamura, K., Suzuki, K., Mori, N., & Takei, N. (2013). Reliability and validity of autism diagnostic interview-revised, Japanese version. *Journal of Autism and Developmental Disorders*, 43(3), 643-662. doi:10.1007/s10803-012-1606-9.

Turygin, N. C., Matson, J. L., Adams, H., & Belva, B. (2013). The effect of DSM-5 criteria on externalizing, internalizing, behavioral and adaptive symptoms in children diagnosed with autism. *Developmental neurorehabilitation*, 16(4), 277-282. doi:10.3109/17518423.2013.769281.

van Heijst, B. F., & Geurts, H. M. (2015). Quality of life in autism across the lifespan: a meta-analysis. *Autism*, 19(2), 158-167. doi:10.1177/1362361313517053.

van Roekel, E., Scholte, R. H., & Didden, R. (2010). Bullying among adolescents with autism spectrum disorders: prevalence and perception. *Journal of Autism and Developmental Disorders*, 40(1), 63-73. doi:10.1007/s10803-009-0832-2.

Ventola, P. E., Kleinman, J., Pandey, J., Barton, M., Allen, S., Green, J., Robins, D., & Fein, D. (2006). Agreement among four diagnostic instruments for autism spectrum disorders in toddlers. *Journal of Autism and Developmental Disorders*, 36(7), 839-847. doi:10.1007/s10803-006-0128-8.

Volkmar, F., Klin, A., Siegel, B., Szatmari, P., Lord, C., Campbell, M., Freeman, B. J., Cicchetti, D. V., Rutter, M., Kline, W., & et al. (1994). Field trial for autistic disorder in DSM-IV. *The American journal of psychiatry*, 151(9), 1361-1367. doi:10.1176/ajp.151.9.1361.

Volkmar, F., Langford Booth, L., McPartland, J., & Wiesner, L. (2014). Clinical Evaluation in Multidisciplinary Settings. In F. Volkmar, R. Paul & S. Rogers (Eds.), *Handbook of Autism and Pervasive Developmental Disorders: Assessment, Interventions, and Policy* (4 ed., Vol. 2, pp. 662-672). Hoboken: Wiley. Retrieved from http://KI.eblib.com/patron/FullRecord.aspx?p=947857.

Volkmar, F., & Reichow, B. (2013). Autism in DSM-5: progress and challenges. *Molecular Autism*, 4(1), 13. doi:10.1186/2040-2392-4-13.

Volkmar, F., Siegel, M., Woodbury-Smith, M., King, B., McCracken, J., State, M., & American Academy of Child and Adolescent Psychiatry Committee on Quality Issues (CQI). (2014). Practice parameter for the assessment and treatment of children and adolescents with autism spectrum disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, 53(2), 237-257. doi:10.1016/j.jaac.2013.10.013.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children - Fourth Edition (WISC-IV) (Swedish version 2007)*. Stockholm, Sweden: Pearson Assessment.

Wechsler, D. (2004). *Wechsler Preschool and Primary Scale of Intelligence - Third Edition (Swedish version 2005)*. Stockholm, Sweden: Pearson Assessment.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale - Fourth Edition (WAIS-IV) (Svedish version 2010)*. Stockholm, Sweden: Pearson Assessment.

Weitlauf, A. S., Gotham, K. O., Vehorn, A. C., & Warren, Z. E. (2014). Brief report: DSM-5 "levels of support:" a comment on discrepant conceptualizations of severity in ASD. *Journal of Autism and Developmental Disorders*, 44(2), 471-476 doi:10.1007/s10803-013-1882-z.

White, S. W., & Roberson-Nay, R. (2009). Anxiety, social deficits, and loneliness in youth with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 39(7), 1006-1013. doi:10.1007/s10803-009-0713-8.

White, S. W., Smith, L. A., & Schry, A. R. (2014). Assessment of global functioning in adolescents with autism spectrum disorders: utility of the Developmental Disability-Child Global Assessment Scale. *Autism*, 18(4), 362-369. doi:10.1177/1362361313481287.

Wiggins, L. D., & Robins, D. L. (2008). Brief report: excluding the ADI-R behavioral domain improves diagnostic agreement in toddlers. *Journal of Autism and Developmental Disorders*, 38(5), 972-976. doi:10.1007/s10803-007-0456-3.

Williamson, S., Craig, J., & Slinger, R. (2008). Exploring the relationship between measures of self-esteem and psychological adjustment among adolescents with Asperger syndrome. *Autism*, 12(4), 391-402. doi:10.1177/1362361308091652.

Wing, L., Leekam, S. R., Libby, S. J., Gould, J., & Larcombe, M. (2002). The Diagnostic Interview for Social and Communication Disorders: background, inter-rater reliability and clinical use. *Journal of Child Psychology and Psychiatry*, 43(3), 307-325. doi:10.1111/1469-7610.00023.

Wittchen, H. U. (1994). Reliability and validity studies of the WHO--Composite International Diagnostic Interview (CIDI): a critical review. *Journal of Psychiatric Research*, 28(1), 57-84. doi:10.1016/0022-3956(94)90036-1.

Wolff, J. J., Gu, H., Gerig, G., Elison, J. T., Styner, M., Gouttard, S., Botteron, K. N., Dager, S. R., Dawson, G., Estes, A. M., Evans, A. C., Hazlett, H. C., Kostopoulos, P., McKinstry, R. C., Paterson, S. J., Schultz, R. T., Zwaigenbaum, L., & Piven, J. (2012). Differences in White Matter Fiber Tract Development Present From 6 to 24 Months in Infants With Autism. *The American journal of psychiatry*, 169(6), 589-600. doi:10.1176/appi.ajp.2011.11091447.

World Health Organization [WHO]. (1992). *The ICD-10 classification of mental and behavioural disorders : clinical descriptions and diagnostic guidelines*. Geneva: Geneva : World Health Organization.

World Health Organization [WHO]. (1993). *The ICD-10 classification of mental and behavioural disorders : diagnostic criteria for research*. Geneva: World Health Organization.

World Health Organization [WHO]. (1994). *Composite International Diagnostic Interview (CIDI): researcher's manual*: World Health Organization.

World Health Organization [WHO]. (2001). *The International Classfication of Functioning, Disability and Health: ICF*. Geneva: World Health Organization.

World Health Organization [WHO]. (2007). *International Classification of Functioning, Disability and Health for Children and Youth (ICF-CY)*. Geneva: World Health Organization.

Üstün, T. B., Chatterji, S., Kostanjsek, N., Rehm, J., Kennedy, C., Epping-Jordan, J., Saxena, S., von Korff, M., & Pull, C. (2010). Developing the World Health Organization Disability Assessment Schedule 2.0. *Bulletin of the World Health Organization*, 88(11), 815-823 doi:10.2471/blt.09.067231.

Zablotsky, B., Bradshaw, C. P., Anderson, C. M., & Law, P. (2014). Risk factors for bullying among children with autism spectrum disorders. *Autism*, 18(4), 419-427. doi:10.1177/1362361313477920.

Zavaleta-Ramirez, P., Nafate-Lopez, O., Villarreal-Valdes, G., Ulloa-Flores, R. E., & Albores-Gallo, L. (2014). K-SADS-PL-2009/Autism spectrum disorders (ASD) inter-rater reliability. *Salud Mental*, 37(6), 461-466.

9 APPENDIX

Diagnostic criteria for Childhood autism according to ICD-10, Autistic Disorder according to DSM-IV and Autism Spectrum Disorder according to DSM-5

ICD-10	DSM-IV/-TR	DSM-5
F84 Childhood autism	**Diagnostic criteria for 299.00 Autistic Disorder**	**Autism Spectrum Disorder 299.00 (F84.0)**
		Diagnostic Criteria
A. Abnormal or impaired development is evident before the age of 3 years in at least one of the following areas: (1) receptive or expressive language as used in social communication; (2) the development of selective social attachments or of reciprocal social interaction; (3) functional or symbolic play.		
B. A total of at least six symptoms from (1), (2), and (3) must be present, with at least two from (1) and at least one from each of (2) and (3). (1) Qualitative abnormalities in reciprocal social interaction are manifest in at least two of the following areas: (a) failure adequately to use eye-to-eye gaze, facial expression, body posture, and gesture to regulate social interaction; (b) failure to develop (in a manner appropriate to mental age, and despite ample opportunities) peer relationships that involve a mutual sharing of interests, activities, and emotions; (c) lack of socio-emotional reciprocity as shown by an impaired or deviant response to other people's emotions; or lack of modulation of behaviour according to social context; or a weak integration of social, emotional, and communicative behaviours; (d) lack of spontaneous seeking to share enjoyment,	A. A total of six (or more) items from (1), (2), and (3), with at least two from (1), and one each from (2) and (3): (1) qualitative impairment in social interaction, as manifested by at least two of the following: (a) marked impairment in the use of multiple nonverbal behaviors such as eye-to-eye gaze, facial expression, body postures, and gestures to regulate social interaction (b) failure to develop peer relationships appropriate to developmental level (c) a lack of spontaneous seeking to share enjoyment, interests, or achievements with other people (e.g., by a lack of showing, bringing, or pointing out objects of interest) (d) lack of social or emotional reciprocity	A. Persistent deficits in social communication and social interaction across multiple contexts, as manifested by the following, currently or by history (examples are illustrative, not exhaustive; see text): 1. Deficits in social-emotional reciprocity, ranging, for example, from abnormal social approach and failure of normal back-and-forth conversation; to reduced sharing of interests, emotions, or affect; to failure to initiate or respond to social interactions. 2. Deficits in nonverbal communicative behaviors used for social interaction, ranging, for example, from poorly integrated verbal and nonverbal communication; to abnormalities in eye contact and body language or deficits in understanding and use of gestures: to a total lack of facial expressions and nonverbal communication. 3. Deficits in developing, maintaining, and

interests, or achievements with other people (e.g. a lack of showing, bringing, or pointing out to other people objects of interests to the individual).

(2) Qualitative abnormalities in communication are manifest in at least one of the following areas:
(a) a delay in, or total lack of, development of spoken language that is *not* accompanied by an attempt to compensate through the use of gesture or mime as an alternative mode of communication (often preceded by a lack of communicative babbling);
(b) relative failure to initiate or sustain conversational interchange (at whatever level of language skills is present), in which there is reciprocal responsiveness to the communications of the other person;
(c) stereotyped and repetitive use of language or idiosyncratic use of words or phrases;
(d) lack of varied spontaneous make-believe play or (when young) social imitative play.

(3) Restricted, repetitive, and stereotyped patterns of behaviour, interests, and activities are manifest in at least one of the following areas:
(a) an encompassing preoccupation with one or more stereotyped and restricted patterns of interest that are abnormal in content or focus; or one or more interests that are abnormal in their intensity and circumscribed nature though not in their content or focus;
(b) apparently compulsive adherence to specific, non-functional routines or rituals;

(2) qualitative impairments in communication as manifested by at least one of the following:
(a) delay in, or total lack of, the development of spoken language (not accompanied by an attempt to compensate through alternative modes of communication such as gesture or mime)
(b) in individuals with adequate speech, marked impairment in the ability to initiate or sustain a conversation with others
(c) stereotyped and repetitive use of language or idiosyncratic language
(d) lack of varied, spontaneous make-believe play or social imitative play appropriate to developmental level

(3) restricted repetitive and stereotyped patterns of behavior, interests, and activities, as manifested by at least one of the following:
(a) encompassing preoccupation with one or more stereotyped and restricted patterns of interest that is abnormal either in intensity or focus
(b) apparently inflexible adherence to specific, nonfunctional routines or rituals
(c) stereotyped and repetitive motor mannerisms

understanding relationships, ranging, for example, from difficulties adjusting behavior to suit various social contexts; to difficulties in sharing imaginative play or in making friends; to absence of interest in peers.

Specify current severity:
Severity is based on social communication impairments and restricted, repetitive patterns of behavior (see Table 2).

B. Restricted, repetitive patterns of behavior, interests, or activities, as manifested by at least two of the following, currently or by history (examples are illustrative, not exhaustive; see text):
1. Stereotyped or repetitive motor movements, use of objects, or speech (e.g., simple motor stereotypies, lining up toys or flipping objects, echolalia, idiosyncratic phrases).
2. Insistence on sameness, inflexible adherence to

(c) stereotyped and repetitive motor mannerisms that involve either hand or finger flapping or twisting, or complex whole body movements;

(d) preoccupations with part-objects or non-functional elements of play materials (such as their odour, the feel of their surface, or the noise or vibration that they generate).

C. The clinical picture is not attributable to the other varieties of pervasive developmental disorder: specific developmental disorder of receptive language (F80.2) with secondary socio-emotional problems; reactive attachment disorder (F94.1) or disinhibited attachment disorder (F94.2); mental retardation (F70-F72) with some associated emotional or behavioural disorder; schizophrenia (F20.–) of unusually early onset; and Rett's syndrome (F84.2).

(e.g., hand or finger flapping or twisting, or complex whole-body movements)

(d) persistent preoccupation with parts of objects

---

routines, or ritualized patterns of verbal or nonverbal behavior (e.g., extreme distress at small changes, difficulties with transitions, rigid thinking patterns, greeting rituals, need to take same route or eat same food every day).

3. Highly restricted, fixated interests that are abnormal in intensity or focus (e.g., strong attachment to or preoccupation with unusual objects, excessively circumscribed or perseverative interests).

4. Hyper- or hyporeactivity to sensory input or unusual interest in sensory aspects of the environment (e.g., apparent indifference to pain/temperature, adverse response to specific sounds or textures, excessive smelling or touching of objects, visual fascination with lights or movement).

Specify current severity:

Severity is based on social communication impairments and restricted, repetitive patterns of behavior (see Table 2).

B. Delays or abnormal functioning in at least one of the following areas, with onset prior to age 3 years: (1) social interaction, (2) language as used in social communication, or (3) symbolic or imaginative play.

C. The disturbance is not better accounted for by Rett's Disorder or Childhood Disintegrative Disorder.

C. Symptoms must be present in the early developmental period (but may not become fully manifest until social demands exceed limited capacities, or may be masked by learned strategies in later life).

D. Symptoms cause clinically significant impairment in social, occupational, or other important areas of current functioning.

E. These disturbances are not better explained by intellectual disability (intellectual developmental disorder) or global developmental delay. Intellectual disability and autism spectrum disorder frequently co-occur; to make comorbid diagnoses of autism spectrum disorder and intellectual disability, social communication should be below that expected for general developmental level.

Note: Individuals with a well-established DSM-IV diagnosis of autistic disorder, Asperger's disorder, or pervasive developmental disorder not otherwise specified should be given the diagnosis of autism spectrum disorder. Individuals who have marked deficits in social communication, but whose symptoms do not otherwise meet criteria for autism spectrum disorder, should be evaluated for social (pragmatic) communication disorder.

Specify if;

With or without accompanying intellectual impairment

With or without accompanying language impairment

Associated with a known medical or genetic condition or environmental factor (Coding note: Use additional code to identify the associated medical or genetic condition.)

Associated with another neurodevelopmental, mental, or behavioral disorder (Coding note: Use additional code[s] to identify the associated neurodevelopmental, mental, or behavioral disorder[s].)

With catatonia (refer to the criteria for catatonia associated with another mental disorder, pp. 119-120, for definition) **(Coding note:** Use additional code 293.89 [F06.1] catatonia associated with autism spectrum disorder to indicate the presence of the comorbid catatonia.)

10 ERRATUM

Unfortunately, in the results section of paper III, p. 193, last part of the last paragraph there is an error. The text should be as follows (changes in bold):

FP (NS children falsely classified as ASD) were younger than TN (NS children correctly classified as NS) on the ADI-R (F(3, 250) = 8.22, post hoc: **T**N > **F**P, $p < .001$, $d = .90$), while FN (children with ASD falsely classified as NS) had higher Vineland-II scores than the TP (children with ASD correctly classified) (F(3, 249) = 34.27, post hoc: **F**N > **T**P, $p < .001$, $d = .83$).