Title

Applications of Item Response Theory to Clinical ADHD Research: Analysis of the Hierarchical Structure of ADHD Symptoms and Increased Precision of Treatment Effect Estimation

Author

Sturm, Alexandra Noelle

Publication Date

2016

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Applications of Item Response Theory to Clinical ADHD Research: Analysis of the Hierarchical Structure of ADHD Symptoms and Increased Precision of Treatment Effect Estimation

A dissertation submitted in partial satisfaction of the

requirements for the degree of Doctor of Philosophy

in Education

by

Alexandra Noelle Sturm

2016

ABSTRACT OF THE DISSERTATION

Applications of Item Response Theory to Clinical ADHD Research: Analysis of the Hierarchical Structure of ADHD Symptoms and Increased Precision of Treatment Effect Estimation

by

Alexandra Noelle Sturm

Doctor of Philosophy in Education

University of California, Los Angeles, 2016

Professor Connie L. Kasari, Chair

Attention-deficit/hyperactivity disorder (ADHD) is a neurodevelopmental disorder with childhood onset that confers greater risk for many negative outcomes including future psychopathology, increased risk for substance abuse, poor self-esteem, and poor social functioning. Identification of targeted treatment approaches requires not only accurate measurement of the construct under study, but also an analytic method that can appropriately identify the effect of treatment, while accounting for the noise of measurement error. This study aimed to address the evident gap in our ability to predict an individual's response to treatment by first identifying a best-fitting model for ADHD symptoms and then using this best fitting model to estimate treatment effect with increased precision.

A sample of 1,612 children and adolescents ages 6 to 17 was used to test the best-fitting model for ADHD. Results from the confirmatory model suggest that a modified bifactor model had better fit (BIC = 47902.44, $M_2(1346) = 2944.68$, RMSEA = .03) compared to the unidimensional and correlated factor models and best minimized mean dimension differences across gender and age. In this modified bifactor model, impulsivity items loaded only on the primary dimension, while inattention and hyperactivity items loaded on both the primary dimension and two separate specific dimensions.

In the analysis of treatment effect, two randomized controlled clinical trials of pharmacotherapy for children and adolescents with ADHD were evaluated; DMPH compared to DMPH + guanfacine and atomoxetine compared to placebo. Item parameters generated in study one were used to estimate the treatment models. In the IRT analyses, atomoxetine produced significant reductions in the inattention and the general dimension, however there was more variability in children's response to atomoxetine as measured by the general dimension. When comparing DMPH and combination treatment, combination treatment was superior in the treatment of the general dimension. Examining variability in average scores provided by the IRT analysis, it appeared that combination treatment was more effective in consistently reducing symptoms across children on the inattention dimension, while DMPH was more effective in consistently reducing symptoms across children on the general dimension. Therefore, creating conditionally independent dimensions clearly allowed for more precise modeling of treatment effect. The field of psychiatry, and more broadly treatment research, could benefit substantially from continued use of IRT models.

The dissertation of Alexandra Noelle Sturm is approved.

James T. McCracken

Li Cai

Michael H. Seltzer

Connie L. Kasari, Committee Chair

University of California, Los Angeles

2016

TABLE OF CONTENTS

1. PREFACE 1
2. CHAPTER 1: Using IRT to identify the hierarchical structure of ADHD items and invariance across age and gender 3
2.1. Introduction 3
Measuring the Latent Trait of ADHD 4
Assumptions of IRT 6
Factor Structure of ADHD Items 6
Considerations for Model Selection 8
ADHD and Age 9
Gender Differences in ADHD 9
IRT and ADHD Rating Scales 10
2.2. Method 12
Data Acquisition 12
Participants 12
Measures 13
ADHD-RS-IV. 13
Statistical Analysis 14
Differential Item Functioning 17
2.3. Results 19
Descriptive Statistics 19
ICC 20
Differential Item Functioning 21
Model Fit 21
Restricted Bifactor Between Group 22
Cross-validation 24
Interpretation of Restricted Bifactor Model Parameters 24
2.4. Discussion 25
The Restricted Bifactor Model and ADHD 26
A New Conceptualization of ADHD Symptoms 27
ADHD Symptom Differences by Age and Gender 28
Conclusion 29
Future Directions 29
3. CHAPTER 2: The use of item response theory for the evaluation of medication effects in clinical trials of pediatric ADHD: Increasing precision of measurement. 39
3.1. Introduction 39
Pharmacotherapy 39
Stimulant pharmacotherapy. 40
Non-stimulant pharmacotherapy. 40
Therapeutic Effects on ADHD Symptom Dimensions 41
Using IRT to Model Treatment Effect 42
3.2. Methods 44
Study 1: Guanfacine and DMPH 44
Participants. 44
Procedure. 45
Study 2: Atomoxetine 46

Measures 47
ADHD-RS-IV. 47
Statistical Analyses 47
Descriptives 47
ANOVA. 48
IRT. 48
3.3. Results 50
Study 1: Atomoxetine 50
Participants. 50
ANOVA. 51
IRT. 51
IRT: Consistency of treatment effect. 52
Study 2: TRECC 53
Participants 53
ANOVA. 53
IRT. 54
IRT: Consistency of treatment effect. 54
Comparing Atomoxetine and DMPH versus COMBO: Summary 55
3.4. Discussion 56
Multidimensionality of ADHD and Estimates of Treatment Effect 56
Treatment Findings from IRT Analysis 57
Atomoxetine 57
DMPH and Guanfacine combination treatment. 58
Treatment of hyperactivity dimension. 58
Future Directions 59
Conclusion 60
REFRENCES 67

LIST OF FIGURES

Figure 2-1. Graded model trace lines for two items with k = 4 response categories. Each line represents the corresponding probability of endorsing the kth category given θ (a unidimensional latent trait). 31
Figure 3-1. Diagram of the multigroup longitudinal restricted bifactor model. 61
Figure 3-2. Study participant EAP scores for the general, inattention, and hyperactivity subdimensions of each treatment group for baseline and treatment endpoint. Each chart shows the variability in treatment effect across participants. 66

LIST OF TABLES

Table 2-1. 32
Table 2-2. 33
Table 2-3. 34
Table 2-4. 35
Table 2-5. 36
Table 2-6. 37
Table 2-7. 38
Table 3-1. 62
Table 3-2 63
Table 3-3. 64
Table 3-4. 65

ACKNOWLEDGEMENTS

The work presented here was made possible by the contributions, help, and support of many people. Thank you to the members of my committee, Connie Kasari, Ph.D. for her feedback and support, James T. McCracken, M.D. for his expertise and encouragement, Li Cai, Ph.D. for emphasizing the importance of applied work in quantitative methods, and Michael Seltzer, Ph.D. In addition, thank you to Megan Kuhfeld who went out of her way to help me troubleshoot, brainstorm, and theorize, to my family for their tireless support and to my husband, Ed Sturm, for his voice of reason, love, and formatting prowess. Most importantly, the author would like to thank the children and families who participated in the clinical trials used for the present investigation. Without them and their commitment to research, this work would not have been possible.

Alexandra Noelle Sturm

Education

University of California, Los Angeles 09/2011-06/2013
M.A.: Psychological Studies in Education

Massachusetts Institute of Technology 09/2004-06/2008
B.S.: Brain and Cognitive Sciences

Teaching Experience

Teaching Assistant Winter 2016
University of California, Los Angeles Summer 2015
Course title: ED132: Autism: Mind, Brain, and Education Winter 2015
Instructor: Connie Kasari, Ph.D.

Publications

Sturm, A. N., Rozenman, M., Piacentini, J. C., McGough, J.J., Loo, S.K., & McCracken, J.T. (*under review).* The effect of neurocognitive function on math achievement in pediatric ADHD: Moderating influences of anxious perfectionism and gender.

Sayer, G., McGough, J. J., Levitt, J., Cowen, J., **Sturm, A.**, Castelo, E., & McCracken, J. T. (*under review*). Acute and long-term cardiovascular effects of stimulant, guanfacine, and combination therapy for attention-deficit/hyperactivity disorder.

Bilder, R. M., Loo, S. K., McGough, J. J., Whelan, F., Helleman, G., Sugar, C., Del'Homme, M., **Sturm, A.,** Cowen, J., Hanada, G., & McCracken, J. T. (2016). Cognitive effects of stimulant, guanfacine, and combined treatment in child and adolescent attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child & Adolescent Psychiatry.*

Loo, S. K., Bilder, R., Cho, A., **Sturm, A.**, Cowen, J., Welker, P. W., Levitt, J., Del'Homme, M., Piacentini, J., McGough, J. J., & McCracken, J. T. (2016). Effects of d-methylphenidate, guanfacine, and their combination on EEG resting state cortical activation in ADHD. *Journal of the American Academy of Child & Adolescent Psychiatry.*

McCracken, J. T., McGough, J. J., Loo, S., Levitt, J., Del'Homme, M., Cowen, J., **Sturm, A.**, Whelan, F., Helleman, G., Sugar, C., & Bilder, R. (*under review*). Combined stimulant and guanfacine administration in attention-deficit hyperactivity disorder (ADHD): A controlled, comparative study.

Sturm, A. N. & Chang, S. W. (2015). Pediatric anxiety: A neurocognitive review. In J. Mohlman, T. Deckersbach, & A. Weissman (Eds.), *From Symptom to Synapse: A Neurocognitive Perspective* (45-74). New York, NY: Routledge, Taylor & Francis Group.

McGough, J. J., Loo, S. K., **Sturm, A.**, Cowen, J., Leuchter, A. F., & Cook, I. A. (2015). An eight-week, open-trial, pilot feasibility study of trigeminal nerve stimulation in youth with attention-deficit/hyperactivity disorder. *Brain Stimulation, 8*(2), pp. 299-304.

McGough, J. J., McCracken, J. T., Cho, A. L., Castelo, E., **Sturm, A.**, Cowen, J., Piacentini, J., & Loo, S. K. (2013). A potential electroencephalography and cognitive biosignature for the child behavior checklist-dysregulation profile. *Journal of the American Academy of Child and Adolescent Psychiatry, 52*(11), 1173-1182.

Conference Presentations

Cai, L., Harrell, L., & **Sturm, A. N.** (2015). *The role of item response theory in assessment and evaluation studies*. ICSA Applied Statistics Symposium, Fort Collins, CO.

Sturm, A. N., McGough, J. J., Cowen, J. C., McCracken, J. T. (2014). *Emotional lability in children with ADHD: stability and comparative effects of Guanfacine, stimulant, and combined treatment*. Poster presented at the American Academy of Child & Adolescent Psychiatry Annual Meeting, San Diego, CA.

McGough, J., Loo, S. K., **Sturm, A.**, Cowen, J., Leuchter, A., & Cook, I. (2013). *A Pilot Study of Trigeminal Nerve Stimulation for ADHD.* American Academy of Child and Adolescent Psychiatry 2013 Meeting, Orlando, FL.

Huston-Carico, A., Davis, A., Del'Homme, M., Bilder, R., McCracken, J. T, & Loo, S. (2010, June). *The Effect of Reading Disability on Executive Function in ADHD.* Consortium for Neuropsychiatric Phenomics 2010 Meeting, Los Angeles, CA.

Talcott, T., **Sturm, A. N.**, Del'Homme, M. A., & McCracken, J. T. (2015). *Family cohesion as a predictor of pharmacologic treatment response in children with ADHD.* UCLA Psychology Undergraduate Research Conference Meeting, Los Angeles, CA.

Torres, J., **Sturm, A. N.**, Del'Homme, M. A., & McCracken, J. T. (2015). *The relationship between problem behaviors, parental concern and family and community SES in children with ADHD.* UCLA Psychology Undergraduate Research Conference Meeting, Los Angeles, CA.

Meagher, S., **Sturm, A. N.**, Del'Homme, M. A., & McCracken, J. T. (2013). *Parental Perceived Locus of Control in Children with ADHD.* UCLA Psychology Undergraduate Research Conference Meeting, Los Angeles, CA.

Fellowships/Awards/Honors

Graduate Research Mentorship award	UCLA 2014-2015
Graduate Summer Research Mentorship award	UCLA 2013
Graduate Summer Research Mentorship award	UCLA 2012

Positions

Graduate Student Researcher at the UCLA Semel Institute 08/2011-present
Los Angeles, CA
Clinical trials for the treatment of attention-deficit/hyperactivity disorder and autism spectrum disorders.
Advisors: James McCracken, M.D. & James McGough, M.D.

Manual co-author and group facilitator 04/2015-present
thinkSMART program, UCLA
(*Successful Management of Approach to Responsibilities and Tasks*)

Staff Research Associate at the UCLA Semel Institute 07/2008-08/2011
Los Angeles, CA
Center for Intervention Development and Applied Research, Translational Research to Enhance Cognitive Control
Supervisors: Robert Bilder, Ph.D., Sandra Loo, Ph.D., Melissa Del'homme, Ph.D.

1. PREFACE

Attention-deficit/hyperactivity disorder is an externalizing neurodevelopmental disorder that affects between 5 and 10% of children and adolescents, as many as three youth in every classroom (Merikangas et al., 2010; Polanczyk, de Lima, Horta, Biederman, & Rohde, 2007). A number of rating scales exist for the purpose of measuring symptoms of ADHD for children and adolescence including the CPRS (Conners' Parent Rating Scale-Revised; C. Keith Conners, Sitarenios, Parker, & Epstein, 1998), SNAP-IV (Swanson, Nolan, and Pelham–IV Questionnaire; Swanson, 1992), and the ADHD-RS-IV (ADHD Rating Scale IV; DuPaul, Power, Anastopoulos, & Reid, 1998). Rating scales serve as a necessary tool for clinicians who wish to determine presence and severity of ADHD symptoms, in addition to monitoring response to treatment (Conners, 1998). Because the best method for detecting the effectiveness of a treatment is through perceived changes in symptoms, accurate measurement of symptomatology is essential. However, as advances in quantitative methodology have improved measurement in domains such as physical functioning, emotional distress, and pain, clinical research in ADHD has lagged behind.

Item response theory (IRT), a latent variable model, presents an opportunity to improve the way we measure baseline clinical symptoms of ADHD in addition to treatment response. The National Institutes of Health and the Patient-Reported Outcomes Measurement Information System (PROMIS) Cooperative Group (Cella et al., 2007) have already taken advantage of the benefits of IRT to produce improved scales for a number of health conditions. To maximize the potential of IRT application in ADHD research, we must also capitalize on the increasing availability of data previously collected by other investigators and companies that have a stake in the advancement of research and treatment.

Beyond improving scales, the effect of treatment can also be modeled in an IRT framework. But, latent variable models such as IRT with many model parameters require large samples for stable and precise estimation. The advent of big data has catalyzed a change in the climate surrounding data use and sharing. From the mandate for data sharing of NIH-funded projects (National Institutes of Health, 2003) to the creation of "transparency" departments at pharmaceutical companies, the last few years are making it increasing possible to take advantage of the availability of data. Now, many samples are available for public consumption and investigators can use more complex models to answer more interesting questions.

This study is presented in two parts. Study 1 determines best model fit and evaluates the psychometric properties of ADHD symptoms. Using the item parameter estimates generated in study 1, study 2 illustrates how investigators can model change in ADHD symptoms as a function of treatment within the IRT framework for improved precision of estimation.

2. CHAPTER 1: Using IRT to identify the hierarchical structure of ADHD items and invariance across age and gender

2.1. Introduction

Measurement of attention-deficit/hyperactivity disorder (ADHD) symptoms is crucial to diagnosis and standardized assessment of symptom change. Most frequently, ADHD symptoms are assessed using rating scales that contain items relevant to the diagnostic criteria. There are a number of measurement theories that aim to identify and resolve pervasive measurement problems. Thus far, all existing scales of ADHD symptoms have been validated using exclusively a classical test theory (CTT) approach. However classical test theory has several shortcomings. Namely, measures with different item content are not directly comparable, all scores in a population are assumed to be measured with equal precision, person ability and item difficulty cannot be separately modeled, and norms are often population dependent (Embretson & Reise, 2000). Therefore, our assessment of a child's ADHD symptoms is limited by the restrictions of classical test theory.

Item response theory (IRT) is a psychometric theory that is becoming increasingly useful to psychological measurement and can address many of the shortcomings of CTT (Embretson & Reise, 2000). IRT assumes that an underlying (latent) trait can explain an individual's responses on a measure, assessment, or test. Through IRT, characterization of symptoms of a disorder can be explored, and unique information provided by IRT analyses can better inform the quality and precision with which we measure disorders. There is unique benefit in using IRT to examine the psychometric properties of a scale *if* the dimensionality of symptoms is modeled appropriately. ADHD is currently assessed as a bi-dimensional construct. Inattention and hyperactivity/impulsivity are two distinct symptom clusters that are assumed unrelated during

assessment. However, there is significant empirical evidence that suggests ADHD is, in fact, not well characterized by a bi-dimensional model and a multidimensional model can better account for residual dependence between symptoms (Toplak et al., 2012; Ullebø, Breivik, Gillberg, Lundervold, & Posserud, 2012) . Investigations that have used IRT thus far have failed to model ADHD symptoms in this established multidimensional structure. Therefore, there is an outstanding need to model ADHD symptoms in the IRT framework using an appropriate multidimensional model so that conclusions can be drawn regarding the aggregation, and relative informativeness, of symptoms.

This study aims to (1) use IRT to establish the best-fitting multidimensional structure of the 18 ADHD-RS-IV items in children and adolescents referred for ADHD diagnosis and treatment, (2) evaluate the item parameters of the IRT model including discrimination, thresholds, and the unique information provided by items in order to draw conclusions about the organization of ADHD symptoms, and (3) generate IRT item parameter estimates that will be used in the analyses of treatment effect where small sample sizes limit the ability to estimate both item parameters, and change over time.

Measuring the Latent Trait of ADHD

Traditionally, summed scores are used to characterize presence and frequency of ADHD symptoms. In IRT, individuals fall along a continuum that reflects the latent trait of ADHD, which represents severity of ADHD symptoms as measured by a specific instrument (e.g. ADHD-RS-IV). In this investigation, severity of the latent trait refers to frequency of ADHD symptoms and not severity of functional impairment. All individuals at the same severity level on the ADHD latent trait have the same probability of a particular response (e.g. *often*) to a specific item (e.g. *talks excessively*). When the possible responses to an item are ordinal, the

graded response model (GRM) for IRT is used (Samejima, 1969). In a unidimensional graded IRT model, each item is described by two types of parameters, a slope/discrimination parameter (α_i) and threshold parameters (β_{ij}). Discrimination parameters represent the relationship between an item and the latent trait and are analogous to a factor loading in factor analysis. Higher discrimination parameters indicate stronger relationships between an item and the latent trait.

In a GRM, for every item (*i*) there are (*k* - 1) thresholds where *k* is the total number of response categories per item in the scale. Each of the threshold parameters in the GRM represents the latent symptom severity the individual must have in order to respond above threshold *k* - 1 with .50 probability. It should be noted that for the GRM, these response thresholds are ordered along the latent trait. Therefore, it is ideal that incrementally higher thresholds are well spaced along the latent trait. When response thresholds are well spaced, an individual who falls high on the latent trait has a high probability of endorsing the most severe response category while an individual low on the latent trait may be most likely to endorse only the lowest response category. For example, Figure 2-1 contains a sample trace line plots for the ADHD items *loses things* (Figure 2-1a) and *on the go* (Figure 2-1b). Each of 4 possible response categories (ranging from *never* to *often)* has its own curve. Comparing the most severe response, *often,* for both items can reveal how differences in the location of response thresholds affect our interpretation of items. For the item *loses things,* a child one standard deviation below the mean is still most likely to endorse the most severe response category, *often.* In contrast, for the item *on the go,* a child who is three standard deviations above the mean is most likely to endorse the most severe response, *often.* This example illustrates how items can provide information at different levels of the latent trait. The combination of discrimination and location of discrimination (thresholds) is unique to IRT and provides valuable information when evaluating

a scale. When creating and evaluating a measure, we hope that items cover the latent trait well, giving us the ability to tease apart incremental increases in the latent trait of ADHD.

Assumptions of IRT

In order to conduct IRT analysis for a measure using a given sample, three primary assumptions must be met: appropriate dimensionality (e.g. number of ADHD traits), local independence, and a correct specification of latent variable distribution in the population (Embretson & Reise, 2000). *First,* the assumption of appropriate dimensionality states that the correct number of traits are measured in order to sufficiently model dependence between the items. *Second,* when items are locally independent, it means that the items are unrelated after controlling for the latent trait. Therefore, it is essential that a conclusion is drawn regarding how to best model the multidimensionality of ADHD. Unmodeled residual dependence between items may result in biased parameter estimates (Chen & Thissen, 1997), and overestimation of IRT item information and test reliability (Sireci, Thissen, & Wainer, 1991).

Factor Structure of ADHD Items

Current diagnostic conceptualizations of ADHD as consisting of two subtypes – inattention and hyperactivity/impulsivity – are not substantiated by factor analytic studies that explored alternative structures of ADHD symptoms. The unidimensional, two- and three-factor models, and the bifactor model have been compared in both clinical and non-clinical samples. The bifactor model has consistently proven to provide superior fit across measures and raters in many large samples. A bifactor model consists of one general factor and several specific factors. In the case of ADHD, the general factor accounts for the latent dimension that underlies all of the symptoms. The specific factors (e.g. inattention and hyperactivity/impulsivity) then account for the residual dependence, or the residual relationship, between items after accounting for the

general factor. In the bifactor model, the general and specific factors are uncorrelated, yielding mutually orthogonal factors. In other words, each factor accounts for unique shared variability among symptoms (Gibbons & Hedeker, 1992).

Furthermore, there is also external validation for the bifactor model. The conditionally independent general and specific factors (Gibbons & Hedeker, 1992) have been found to uniquely predict several academic (Willoughby, Blanton, & Investigators, 2015), cognitive (Smith, Tamm, Hughes, & Bernstein, 2013) and social-emotional (Willoughby et al., 2015) outcomes. In addition, the general and specific factors show differential association with constructs, such as cognitive control, while ADHD symptom totals do not (Martel, Roberts, Gremillion, von Eye, & Nigg, 2011). This suggests the possible utility in parsing unique variability in symptoms of ADHD into a general factor and precise factors representing inattention and hyperactivity/impulsivity.

However, evidence suggests that a restricted bifactor model may be more suitable when examining the dimensionality of ADHD. A restricted bifactor model is useful when a construct is generally unidimensional but excess covariance exists between some clusters of items. Items load on a specific factor *only when necessary* (Stucky & Edelen, 2015). In nearly every study that tested the two specific factor bifactor model, factor loadings for the hyperactivity/impulsivity factor were opposing for hyperactivity items and impulsivity items. When the hyperactivity items (e.g. *driven by a motor*, *runs or climbs excessively*, *fidgets*) had stronger loadings on the secondary factor, the impulsivity items (e.g. *talks excessively*, *difficulty awaiting turn*, *interrupts or intrudes on others*) had either negative or negligible loadings and vice versa (Gibbins, Toplak, Flora, Weiss, & Tannock, 2012; Gomez, Vance, & Gomez, 2013; Martel, Eye, & Nigg, 2012; Normand, Flora, Toplak, & Tannock, 2012; Smith et al., 2013;

Toplak et al., 2009; Toplak, West, & Stanovich, 2012; Wagner et al., 2016). Therefore, the factor is primarily defined by hyperactivity, and the impulsivity items are then inversely related to that factor. Clinically, this suggests that there are possibly several underlying phenotypes related to hyperactivity and impulsivity. One phenotype that represents the underlying behavior common to hyperactivity and impulsivity, and two remaining phenotypes that are unique to the symptoms of hyperactivity and impulsivity. The bifactor model with three specific factors, which would isolate the hyperactivity and impulsivity items into two separate secondary factors, was sometimes inestimable (Smith et al., 2013; Ullebø et al., 2012). Discrepant loadings on a secondary factor suggest that a restricted bifactor model may be more appropriate. Ullebø et al. (2012) found excellent and parsimonious fit with a modified bifactor, specifying only specific factors for the inattention and impulsivity items, while the hyperactivity items loaded exclusively on the general dimension.

Considerations for Model Selection

It is critical for both scientific investigation and clinical assessment and intervention that our ability to measure ADHD is consistent and precise across key demographic characteristics. IRT can help to determine if probability of a specific response to a particular item, or the location at which the item is able to best discriminate between individuals at a specific severity level differs by gender or age. Generally speaking, we must be able to answer the question: do ADHD symptoms manifest differently across gender or age groups? Overall mean differences are of concern, but if the probability of symptom endorsement at a specific trait level varies as a function of age or gender, greater concern should be raised about the use of uniform diagnostic criteria across gender.

ADHD and Age

Overall, a diagnosis of ADHD is highly persistent, with rates declining over time differentially by subtype (Willcutt, 2012). An ADHD diagnosis as early as preschool is relatively stable, with between 67% and 89% of children retaining their diagnosis through elementary school (Lahey, Pelham, Loney, Lee, & Willcutt, 2005; Riddle et al., 2013). Subtype and severity both play a role in persistence, as children with ADHD-combined type are most likely to retain an ADHD diagnosis (Lahey et al., 2005) and severity predicts persistence of the disorder into adulthood (Biederman et al., 2006).

Interestingly, the hyperactive subtype is relatively unstable during early developmental years. Around the onset of puberty, symptoms of hyperactivity/impulsivity begin to decline, while symptoms of inattention persist (Biederman, Mick, & Faraone, 2000; Hart, Lahey, Loeber, Applegate, & Frick, 1995; Willcutt et al., 2012). The changing manifestation of inattention and hyperactivity/impulsivity over development requires close attention to possible differences in the probability of specific ADHD item endorsement (differential item functioning).

Gender Differences in ADHD

Estimates range (2:1-9:1; Rucklidge, 2010) in the degree to which boys out-represent girls with ADHD, as do typical patterns of symptom presentation (Willcutt, 2012); however, in general prevalence rates of ADHD for boys are greater than for girls. Girls display fewer symptoms of inattention, hyperactivity, and impulsivity relative to boys (Arnett, Pennington, Willcutt, DeFries, & Olson, 2015; Gershon & Gershon, 2002), although estimates vary for clinic-referred samples and non-referred samples. Studies suggest that greater overt behaviors, including rule-breaking and externalizing behaviors, seen in boys with ADHD (Abikoff et al.,

2002) may explain why boys are more often referred for treatment than girls. In fact, girls are twice as likely to be diagnosed with inattentive type alone (Biederman et al., 2002).

Boys receive more severe ratings on ADHD symptoms than girls, and they also exhibit greater variability in symptom totals. Higher mean ratings and increased variability for males result in more males exceeding the diagnostic threshold (Arnett et al., 2015). Systematic differences in both mean and variance estimates for male and female ADHD symptoms are of concern when we consider the measures that are used for diagnosis. It is possible that, due to measurement, girls who report subthreshold symptoms compared to the population, but require additional services, do not receive a diagnosis. One way to begin addressing this issue would be to identify symptom-level differences by gender. No studies have systematically examined specific *symptom* differences between girls and boys. Exploring gender as a source of differential item functioning (DIF), or different probability of response to an item depending on group membership, is necessary for continued use of ADHD rating scales.

IRT and ADHD Rating Scales

Researchers have recently begun to examine the psychometric properties of ADHD rating scales using IRT. IRT models have been fit to ADHD rating scales for preschoolers (Purpura, Wilson, & Lonigan, 2010) and adolescents (Garcia-Rosales et al., 2015; Gomez, 2008) with varying success. For the most part, ADHD items provide high and non-redundant information at different locations on the latent trait (Garcia-Rosales et al., 2015; Gomez, 2008), indicating that most items are representative of the underlying latent trait of ADHD. While there was some overlap between these three studies in items deemed the poorest performers across the three scales, the items were presented for different, theoretically overlapping reasons (low discrimination in Garcia-Rosales et al., 2015 and Purpura, Wilson and Lonigan, 2010 vs. low

item information in Gomez, 2008). In addition, both Rosales et al. (2014) and Gomez (2008) used parent report of their child's symptoms; however, the use of special instructors as raters who had relatively little exposure to each preschooler's behaviors in Purpura, Wilson and Lonigan (2010) limits our ability to compare and contrast conclusions from each study.

Nonetheless, no studies that employed IRT for ADHD symptoms reported model goodness-of-fit statistics and none modeled the items in the context of a multidimensional model, which we know to be essential for an accurate understanding of ADHD. At its core, IRT is concerned with modeling the relationship between latent traits and responses to items, or the plausibility of a process for item selection (Maydeu-Olivares, 2013). Without model fit as a frame of reference, it is difficult to know how our conceptualization of the process fits what we observe in our data. Furthermore, none of the studies systematically examined measurement invariance through differential item functioning. Evidence exists for differences by gender in the strength of item association with ADHD factors. Confirmatory factor analysis of Australian adolescents revealed measurement invariance of self-reported ADHD symptoms across gender (Gomez et al., 2013). Separate norms for boys and girls for many ADHD scales also support the need to examine DIF (DuPaul et al., 1998) – whether the items function differently for girls and boys, or if girls on average occupy a different mean latent trait score compared to boys.

As described above, the exploration of the 18 ADHD symptoms using IRT is a novel approach that can improve our understanding of symptoms and identify possible differences in probability of symptom endorsement by age and gender. Although researchers have used parent and teacher report for exploration of symptoms, the field has yet to use a measure of ADHD symptoms rated by parents and facilitated by a trained clinician. The use of a trained clinician may improve rigor of assessment and reduce measurement error, as parents may perceive item

questions and ratings differently. In addition, the analyses of different age groups hint at possible differential item functioning, however conclusions based on comparison of findings are somewhat obfuscated by different standards for raters. Measurement invariance in adolescence was found between genders (Gomez et al., 2013) in the CFA framework, however systematic DIF testing within IRT has yet to be explored. Finally, parameter estimates generated in this paper may be used for more complex models of treatment effect.

2.2. Method

Data Acquisition

Data was requested in June 2015 from three pharmaceutical companies and from investigators at the University of California, Los Angeles (UCLA), a total of 20 studies from the four data holders. Five studies were requested from UCLA and the request was approved and data immediately released. Study proposals were prepared for and data from a total of 15 studies were requested from the three pharmaceutical companies. One of the companies responded that despite announcement of their commitment to data sharing and the possibility of data de-identification, the informed consents for the referenced studies did not permit patient-level data sharing. In July of 2015, the data request was approved by the two remaining pharmaceutical companies. One company immediately initiated data de-identification. The first study was received in October of 2015 and the last study received in January of 2016. The investigator has yet to receive data from the third company and recently received an update that data may be released by June of 2016.

Participants

To obtain a sample size sufficiently large for the proposed analyses, baseline data was acquired by this author from six separate industry sponsored trials (study identification numbers:

B4Z-MC-LYAC, B4Z-MC-LYAS, B4Z-MC-LYAQ, B4Z-US-LYBP, B4Z-US-LYCC, B4Z-MC-LYAT) and four investigator-initiated trials (UCLA1 [NCT01714310; U01MH093582]; UCLA2 [NCT01388530]; UCLA3 [R34MH101282, NCT02155608]; UCLA4 [K23MH01966]). The studies included children ranging from 6 to 17 (*n* = 1612*),* all recruited for treatment trials for ADHD and/or ADHD in addition to a comorbid condition.

In each trial, diagnoses were determined using the Kiddie Schedule for Affective Disorders and Schizophrenia for School Aged Children-Present and Lifetime Version (K-SADS-PL; Kaufman et al., 1997), a semistructured diagnostic interview administered by a trained clinician. Due to the purpose of the trials from which this data was acquired, several samples were predominantly comprised of study participants with specific comorbid conditions (e.g. B4Z-US-LYBP, anxiety; B4Z-MC-LYAQ, mixed anxiety and affective; B4Z-MC-LYAS, Tourette disorder or chronic motor tics). For the current investigation, effort was made to select studies for the development and validation samples to maximize clinical and demographic heterogeneity. Comorbidity in each sample population is roughly reflective of diagnostic rates (see Table 2-1.). For the purpose of these analyses, data from all children and adolescents screened for whom an ADHD-RS-IV was completed were included, and it was not necessary that each child met individual study inclusion criteria or diagnostic criteria for ADHD.

Before participating in any of these trials, study participants and parents provided informed assent and consent approved by the site's institutional review board.

Measures

ADHD-RS-IV.

The ADHD Rating Scale–IV (ADHD-RS-IV; DuPaul et al., 1998) is a clinician-rated measure based on the ADHD diagnostic criteria from the DSM-IV-TR (American Psychiatric

Association, 2000) and is widely considered a useful measure for detecting treatment outcome in children and adolescents with ADHD. The ADHD-RS-IV contains nine inattention and nine hyperactivity-impulsivity items that are presented in alternating order. Items include "fails to give close attention to details or makes careless mistakes," "loses things necessary for tasks or activities," and "leaves seat in classroom or in other situations in which remaining seated is expected." Each item is rated on a 4-point Likert scale, where 0 = *never or rarely,* 1 = *sometimes,* 2 = *often,* 3 = *very often,* resulting in a minimum score of 0 and a maximum score of 27 for each subscale. In each of the samples, a parent or legal guardian was asked by a trained clinician to rate the frequency of their child's behavior. At the time of the assessment, children and adolescents were either unmedicated, or rated based on behavior off-medication. For the present investigation, screening ADHD-RS-IV data was used.

Statistical Analysis

Descriptive information was computed for the entire sample and by study (see Table 2-1., Table 2-2., and Table 2-3.). Seven datasets (B4Z-MC-LYAC, B4Z-MC-LYAS, B4Z-MC-LYAQ, B4Z-US-LYBP, B4Z-US-LYCC, UCLA3, UCLA4) were combined to form an exploratory sample (n = 1357) and three datasets (B4Z-MC-LYAT, UCLA1, UCLA2) were combined to form a validation sample (n = 255). Differences between the exploratory and validation samples were examined for ADHD-RS-IV total and subscale raw scores, age and proportion of females. Differences between studies within the exploratory and validation samples were not examined as differences would be expected across trials as all were treatment trials with specific inclusion and exclusion criteria. There were only two missing values in the sample and thus non-random missingness was not a concern and was not explored.

Due to the multilevel nature of the data (participants nested within sites), intraclass correlation coefficients (ICC) were computed for each item to determine if it was necessary to model in a multilevel structure. The ICC represents the between group variation divided by total variation (between plus within), yielding an estimate of the percentage of variation attributable to between-site differences that exists between sites in the sample. ICC values that exceed .10 are often used to determine that the multilevel nature of the data needs to be accounted for in the modeling framework.

Competing IRT models were fit to participants' ADHD-RS-IV item responses. All models were fit using maximum marginal likelihood estimation via the Bock-Aitkin EM algorithm (Bock & Aitkin, 1981) using the software flexMIRT (Cai, 2012) in the exploratory sample (n = 1,357) of children and adolescents. Graded response models (GRM; Samejima, 1969) were used to model the ordinal nature of the ADHD-RS-IV four category responses. In the multidimensional GRM model, each item is characterized by a vector of *slope* or *discrimination (a)* parameters that reflects the strength of association between an item and the factor of interest and three threshold parameters (b_{ic}, corresponding to the boundaries between the four response categories) which show the severity of ADHD at which the probability of responding to a specific category, or higher, is .50 (Samejima, 1969). The cumulative category response probabilities are

$$P(y_{ij} \geq 1 | \theta_{j1}, \theta_{j2}) = \frac{1}{1 + \exp[-(c_{i1} + a_{i1}\theta_{j1} + a_{i2}\theta_{j2})]}$$

$$\vdots \qquad (1)$$

$$P(y_{ij} \geq K - 1 | \theta_{j1}, \theta_{j2}) = \frac{1}{1 + \exp[-(c_{i,k-1} + a_{i1}\theta_{j1} + a_{i2}\theta_{j2})]} \quad .$$

The category response probability is the difference between two adjacent cumulative

probabilities

$$P(Y_{ij} = k|\theta_{j1}, \theta_{j2}) = P(Y_{ij} \geq k|\theta_{j1}, \theta_{j2}) - P(Y_{ij} \geq k+1|\theta_{j1}, \theta_{j2}), \quad (2)$$

where $P(y_{ij} \geq 0|\theta_{j1}, \theta_{j2})$ is equal to 1 and $P(y_{ij} \geq K|\theta_{j1}, \theta_{j2})$ is zero. In all models, we assumed that the latent traits followed a standard normal distribution. However, we checked the violability of this assumption with our final model by approximating the latent trait using an empirical histogram (Mislevy, 1984; Woods, 2007).

First, a unidimensional model was fit to the ADHD-RS-IV items, which assumed that the relationship between items could be explained by a single underlying dimension. Second, a two factor correlated traits model (inattention, hyperactivity/impulsivity) and a three factor correlated traits model (inattention, hyperactivity, impulsivity) were fit, allowing all factors to correlate. Third, four bifactor models were fit. Each model contained one "primary" dimension on which all items loaded, but the number and loadings of the "specific" factors differed by model. Two standard item bifactor models (e.g. Cai, Yang, & Hansen, 2011; Gibbons & Hedeker, 1992) were fit to explore the possibility of two and three specific factors - (1) inattention and hyperactivity/impulsivity or (2) inattention, hyperactivity and impulsivity - and one general dimension that represents general ADHD trait variation. Two restricted bifactor models were fit, with all of the items loading on one general dimension, all inattention items loading also on a specific factor and either hyperactivity or impulsivity items loading on a second specific factor (thus limiting the numbers of parameters freely estimated). Hyperactivity or impulsivity items loaded exclusively on the general dimension in each respective model.

FlexMIRT provides item intercepts, and therefore item thresholds with respect to the general dimension were computed using the following equation for multidimensional models to ease interpretation of results:

$$B_{ic} = \frac{-c_{ic}}{\sqrt{\sum_{k=1}^{m} a_{ik}^2}}$$

Here, c_{ic} is the *c*th intercept parameter for item *i*, a_{ik}^2 is the squared slope parameter for item *i* on dimension *k*, and *m* is the number of dimensions (Reckase, 2009). In multidimensional IRT, the marginal threshold for the general dimension is the location on the general dimension where the probability of endorsing a particular response category is .50 given that all specific dimensions are fixed at zero.

Indices of model fit including log-likelihood, mixed M_2 statistic (Monroe & Cai, 2015), AIC, BIC, local dependence (LD) diagnostic indices (Chen & Thissen, 1997), and consideration of item loadings were used to determine the best fitting model. In addition, to select the final model between models with similar fit, multigroup models were estimated using the entire sample (N = 1612) by age (< 11 and >= 11 years old) and gender. The model that best minimized mean and variance differences between groups was selected as the final model.

Following determination of best model fit, the best model was estimated using a validation sample (n = 255) of children and adolescents. For cross-validation, all item discrimination parameters (a) and intercepts (b) were constrained to the values generated in the development sample, while means and variances were freed to allow for natural between-sample variability.

Differential Item Functioning

DIF was evaluated following the procedures set forth by Hansen et al. (2014) across gender and two age subgroups (6 - 10 vs. 11 - 17) for the inattention items and hyperactivity/impulsivity items in separate, sufficiently unidimensional models. Given the evolving nature of the inattentive and hyperactivity/impulsivity dimensions over time, DIF was

computed for children ages 6 to 10 relative to children ages 11 to 17 due to changes in symptom presentation around puberty (Willcutt, 2012). DIF analyses consisted of three steps; (1) evaluated all items simultaneously for possible DIF, (2) candidate DIF items were evaluated with the remaining items serving as anchors and (3) examination of the severity of DIF. Due to the lack of literature supporting a priori assumptions regarding items that may serve as anchors, the first step therefore identified anchor items and the second step served as the candidate DIF test (Woods, Cai, & Wang, 2013).

The first step consisted of a two-stage Wald χ^2 procedure (Langer, 2008; Woods et al., 2013) to determine candidate items (items that showed initial evidence of DIF), and anchor items (items that showed no initial evidence of DIF). First, all item parameters were constrained equal between groups and means and variances for the focal groups were estimated compared to the reference group with $\mu = 0$ and $\sigma^2 = 1$. Then the means and variances estimated in step 1 were fixed as known and the item parameters were all estimated. Critical *p* values were adjusted for the test statistics using the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) with an overall alpha level of .05 to control for the overall false discovery rate. Items that showed significant group differences according to the corrected alpha level were considered candidate DIF items.

The second step in the DIF procedure was to test the DIF candidate items. Another set of two-group IRT models were evaluated where the parameters for the DIF candidate items were freely estimated in each group and the anchor item parameters were fixed as equal between groups. Again, the Wald χ^2 and Benjamini-Hochberg adjustment were used to test for significant differences between item parameters across groups.

The DIF detection procedure in steps one and two is highly sensitive, and it is possible that statistically significant differences in item parameters may not in fact be substantively significant. To determine the severity of DIF items found to be significant after step two, the weighted area between the expected score curves (wABC) was used (Hansen et al., 2014). Generally, wABC can give us information about the degree of deviation of the expected score curves between groups and is analogous to a measure of effect size. Expected score curves represent the relationship between the underlying trait (ADHD) and the average response, given the possible response categories. The degree of deviation is weighted by the population distribution such that estimated item parameter differences close to the middle of the distribution are weighted more heavily than differences at the tails. A cut-off value of .30 was used to identify items that showed significant DIF.

2.3. Results

Descriptive Statistics

The total sample (N = 1612), mean age 10.68 (SD = 2.56), was 73.82% Caucasian, 12.41% Black or African American, 1.55% Asian, 8.56% Hispanic, and 3.54% other (mixed race, American Indian, Alaska Native). Descriptive statistics were generated for the ADHD-RS-IV inattention and hyperactivity subscales to evaluate the range of data and univariate statistics of normality. Descriptive statistics and sample characteristics were examined for each study (Table 2-2. and Table 2-3.), the total sample (Table 2-1.), and the total sample by gender and by age. Overall, inattention (M = 21.59, SD = 4.61) and hyperactivity (M = 16.91, SD = 6.75) subscales were highly elevated and met or exceeded the 80th percentile cut-off for diagnosis of ADHD as recommended by DuPaul et al. (1998) for all age and gender groups. There were no significant differences between boys and girls on symptoms of inattention (t(729.49) = 1.24, p =

.21), hyperactivity/impulsivity (t(772.60) = 1.95, p =.05), or total symptoms (t(748.53) = 1.95, p =.05). There were no significant differences between participants < 11 and >= 11 on symptoms of inattention (t(1302.57) = 1.61, p <.11), however there were significant differences on symptoms of hyperactivity/impulsivity (t(1272.15) = 9.27, p < .001) and total symptoms (t(1306.86, p < .001). Looking more closely at group means, children under the age of 11 had noticeably more symptoms of hyperactivity/impulsivity (M = 18.17, SD = 6.18) than children 11 and over (M = 15.00, SD = 7.11), while inattention symptoms were relatively equal across groups (<11 y/o: M = 21.73, SD = 4.44; >=11 y/o: M = 21.34, SD = 4.95).

There were no significant differences between the development (n = 1357) and validation (n = 255) samples on symptoms of inattention (t(373.27 = -1.65, p < .10), however there were significant differences on symptoms of hyperactivity/impulsivity (t(348.86) = -2.84, p = .004) and total symptoms (t(368.25) = -2.90, p = .004). The validation sample was significantly younger (M = 10.34, SD = 2.57) than the development sample (M = 10.72, SD = 2.77; t(341.05) = -2.03, p = .04). There were no significant differences in the proportion of girls between samples (X^2(1) = 1.06, p = .30).

ICC

The ICC for each site gives information about the amount of variability attributed to between site differences. In the exploratory sample, site sample sizes ranged from 1 to 46 (M = 16.8, SD = 10.9) and ICCs for the 18 ADHD-RS-IV items ranged from .06 to .14 (M = .09, SD = .02). Site sample sizes in the validation sample ranged from 8 to 40 (M = 23.18, SD = 8.72) and ICCs for the 18 ADHD-RS-IV items ranged from .01 to .26 (M = .07, SD = .06). This implies that most of the variance in ADHD-RS-IV items is within site and thus the variability in measurement was relatively consistent across sites.

Differential Item Functioning

DIF exists when different groups (e.g. gender or age groups) have a different probability of response to an item. None of the items met the wABC criterion for DIF. The item "talks excessively" was on the threshold for possible problematic DIF across gender (wABC = .30), such that males had to occupy a more severe location on the latent trait relative to females in order to endorse the item. However, after examination of the expected score curves, the difference was not deemed sufficiently problematic to warrant removal of the item. Therefore, analysis proceeded with all items.

Model Fit

Results from the confirmatory model suggest that the bifactor model had better fit compared to the unidimensional and correlated factor models. As detailed in Table 2-4, the restricted bifactor model showed the best fit (BIC = 47902.44, $M_2(1346) = 2944.68$, RMSEA = .03) compared to the unidimensional model (BIC = 50263.5, $M_2(1359) = 7834.11$, RMSEA = .06), the two correlated factor model (BIC = 48603.67, $M_2(1358) = 3277.37$, RMSEA = .03), the three correlated factor model (BIC = 48441.67, $M_2(1356) = 3081.87$, RMSEA = .03), and the two specific factor, bifactor model (BIC = 48351.39, $M_2(1341) = 2915.45$, RMSEA = .03). The three specific factor bifactor model did not reach a maximum during EM estimation and therefore results are not reported. In addition, examining local dependence indices (Chen & Thissen, 1997), high positive values (>9; Stucky & Edelen, 2015) revealed that the relationship among inattention items and hyperactivity/impulsivity items respectively were greater than predicted by the unidimensional model. Although the three correlated factor model provided relatively good fit, the correlation between the hyperactivity and impulsivity factor ($r = .85$, $SE = .03$) was too high to warrant consideration of the two factors as distinct. We also checked the

sensitivity of our model to violations of the normal assumption by fitting the model using an empirical histogram (Woods, 2007), and found that for this clinical sample, a normal distribution assumption was sufficient.

Closer examination of the two specific dimension bifactor revealed a problematic hyperactivity/impulsivity specific factor. The hyperactivity items (e.g. *fidgets*, *driven by a motor*) all showed small negative discrimination parameters (-0.15 to -1.06) while the impulsivity items (e.g. talks excessively, blurts out answers) had small positive discrimination parameters (0.40 to 1.26). The three dimension bifactor, a possible solution to the inconsistent loadings, did not estimate properly and therefore two restricted bifactor models were fit: (1) a bifactor with the hyperactivity items loading only on the primary dimension, and two specific dimensions for inattention and impulsivity (bifactor I/I) and (2) a bifactor with the impulsivity items loading on the primary dimension and two specific dimensions for inattention and hyperactivity (bifactor I/H). The two models showed minimal differences in fit indices (see Table 2-5.; bifactor I/I; BIC = 48345.54, $M_2(1346) = 2944.68$; bifactor I/H; BIC = 48371.7, $M_2(1345) = 2949.45$).

Restricted Bifactor Between Group

Due to minimal differences in fit between the two restricted bifactor models, the full sample (development and validation) was used to compare the two restricted models across the age groups (< 11 and >= 11) and gender. A model that showed minimal differences between groups would be advantageous for future evaluation of the construct. Again, indices and model fit were very similar (see Table 2-5.), but differences in group means existed. Mean differences (*MD*) between males and females and younger (< 11) and older (>= 11) children were smaller in the bifactor I/H model compared to the bifactor I/I model.

Males and females showed negligible differences in means for the inattention specific dimension in both the bifactor I/I (*MD* = -0.05, *SE* = .07) and bifactor I/H (*MD* = .09, *SE* = .07) models. In contrast, greater mean differences were observed for the impulsivity specific factor in the bifactor I/I and the hyperactivity factor in the bifactor H/I. Girls were rated .42 (*SE* = .08) standard deviations higher than boys on the impulsivity specific dimension of the bifactor I/I model and boys were rated .52 (*SE* = .08) standard deviations higher than girls on the hyperactivity factor of the bifactor I/H model. In addition, nearly no difference existed in the primary ADHD dimension of the bifactor I/H model (*MD* = .01, *SE* = .06), while in the bifactor I/I model boys were rated .24 standard deviations (*SE* = .06) higher than girls.

More mean differences within restricted bifactor models existed between younger and older children. The bifactor I/H model generally best minimized mean differences in factor scores across age, with exception of the hyperactivity specific dimension for which younger children scored .69 (*SE* = .08) standard deviations higher than older children. In the bifactor I/H model, younger children scored .32 standard deviations (*SE* = .08) higher on the primary dimension than older children and .12 standard deviations (*SE* = .08) lower on the inattention specific dimension. In the bifactor I/I model, younger children scored .26 standard deviations (*SE* = .08) lower than older on the inattention specific dimension and .31 standard deviations (*SE* = .09) higher on the impulsivity specific dimension.

The bifactor model with inattention and hyperactivity specific dimensions possessed good model fit compared to the unidimensional and correlated factor models, and best minimized differences in the general dimension, providing superior fit for the intention of obtaining a general dimension to measure the trait of ADHD. Thus, we tested this restricted bifactor model using the cross-validation sample.

Cross-validation

The restricted bifactor model fit the cross-validation sample reasonably well (BIC = 9457.99, M_2 = 386.89, RMSEA = .06) although clearly defined cut-offs for determining adequacy of model fit in cross-validation have not yet been established. LD indices also reflected adequate model fit as no items exhibited extra-large unexplained residual dependence.

Interpretation of Restricted Bifactor Model Parameters

In IRT, the item intercept parameters for the bifactor model are interpreted as the relationship between the response to an item and the dimension of interest, conditional on all other dimensions on which the item loads (Stucky & Edelen, 2015). To ease interpretation, factor loadings were computed (see Table 2-6.) from the standardized slope parameters (*a*) using equation 3 (Cai, 2010a), where λ_1 is the general dimension loading in factor analytic notation and *D* is the scaling constant 1.7.

$$\lambda_1 = \frac{a_1/D}{\sqrt{1 + \sum(\mathrm{a/D})^2}} \tag{3}$$

Overall, the items loaded highly on both the specific and primary dimensions. The hyperactivity and impulsivity items showed the strongest association with the primary dimension. *Interrupts or intrudes* was the highest loading (λ = .83) impulsivity item and *on the go* was the highest loading hyperactivity item (λ = .74). On the primary dimension, inattention items were on average less discriminating (λ = .39) relative to the level of discrimination on the inattention specific dimension (M_λ = .60). The item *easily distracted* (λ = .55) was the most discriminating inattention item on the primary dimension, followed by *does not listen* (λ = .51) and *difficulty sustaining attention* (λ = .47).

Inattention items were generally discriminating on the specific inattention dimension (M_λ = .60). Of the inattention items, *follow through* was the most discriminating (λ = .70) for the inattention specific dimension and *does not listen* was the least discriminating (λ = .39). The hyperactivity specific dimension was dominated by the item *runs about or climbs* (λ = .60), and the remaining four items were relatively weakly discriminating on the hyperactivity specific dimension (M_λ = .30).

In general, the inattention items can be seen as "easier" compared to the hyperactivity and impulsivity items (i.e., the thresholds of the inattention of items are located around lower levels of severity). For example, the range of severity for the inattention item *does not listen* was -3.17 to .12, which implies that even individuals with moderately low severity on the latent trait were likely to endorse the most severe response category for symptoms of inattention.
In addition, no items had an extremely narrow range of threshold parameters, meaning that each item covered a relatively broad range of the latent trait. However, there were no items that discriminated much more than one standard deviation above the mean severity. Therefore, we infer that this scale is not providing a large degree of information at higher levels of the latent dimensions based on this sample.

2.4. Discussion

The primary aims of this study were to use the modeling precision of IRT to test the factor structure of youth ADHD symptoms and to characterize patterns of item parameters within the context of a multidimensional IRT model. This study revealed that a restricted bifactor model, where only inattention and hyperactivity items loaded on specific factors, provided superior fit to the observed data, with model fit consistent in a cross-validation sample. Using the restricted bifactor model, age and gender differences were most effectively minimized on the

general ADHD dimension and the inattention dimension. The findings in this report provide evidence that separate analysis of the inattention and hyperactivity/impulsivity fail to adequately account for the multidimensionality of ADHD symptoms.

The Restricted Bifactor Model and ADHD

These findings are consistent with other published reports that indicate the bifactor model best fits ADHD symptoms (Toplak et al., 2012; Ullebø et al., 2012). Therefore, prior reports that assumed inattention and hyperactivity/impulsivity were sufficiently independent constructs to be evaluated in separate, unidimensional IRT models were insufficient to make conclusions regarding item (i.e. symptom) parameters (Garcia-Rosales et al., 2015; Gomez, 2008; Purpura et al., 2010). It is essential to model multidimensionality when present because over or under explanation of the relationship between items can not only distort item discrimination parameters (Reise, Morizot, & Hays, 2007), but also artificially inflate estimates of score reliability (Sireci et al., 1991; Thissen, Steinberg, & Mooney, 1989). Item parameters generated by IRT analyses must be evaluated in the context of a model that best fits the construct.

Inattention, hyperactivity, and impulsivity items were not found to be equally discriminating at all locations on the latent trait of ADHD. Generally, all of the ADHD items discriminated well between individuals three standard deviations below the mean to one standard deviation above the mean ADHD severity. Inattention items were discriminating at lower levels of ADHD severity compared to hyperactivity and impulsivity. Given that this was a clinical sample of children and adolescents, most of who were diagnosed with ADHD, the mean severity level may be interpreted as mean severity of a clinical sample. In a community sample of youth, the threshold values of the items would shift (by a simple linear transformation) such that items would be more discriminating at high levels of ADHD severity, thereby covering a more

"severe" portion of the population as would be expected when hoping to determine presence or absence of ADHD. The relationship between items and the primary and specific dimensions (i.e. discrimination parameters) should remain consistent in a different sample. While informing our ability to diagnose, predict outcomes, monitor sensitivity to treatment, and so forth, greater discrimination at more severe levels of ADHD would be incredibly informative to both researchers and clinicians.

A New Conceptualization of ADHD Symptoms

Despite generally good item discrimination on the general and specific dimensions, there were distinct patterns of high and low discrimination that may inform our understanding of ADHD. While many of the inattention items had relatively low discrimination on the general dimension, three inattention items (*sustaining attention*, *listen when spoken to*, and *easily distracted*) appear to drive the relationship between the inattention, hyperactivity, and impulsivity items. Comparing these three items to the remaining inattention items, these three appear to assess attention that requires a more immediate form of cognitive control. For example, a distractor is presented and a child has an immediate impulse to engage with the distractor. In this context, the overlap between the inattention items and the hyperactivity and impulsivity items is clear, all require immediate engagement of cognitive control. Inhibition, or the ability to inhibit a prepotent response, may be the construct that best describes the variability common to these items. The remaining inattention items are more consistently multistep processes for which children may be better able to develop compensatory strategies. *Organization*, *following through on instructions*, *avoids tasks*, etc., do not require immediate impulse control, rather sustained attention and repeated decisions that may result in endorsement of the symptom. *Following*

through on instructions, for example, requires successful completion of a series of steps which must be maintained in working memory, likely planned, and subsequently executed.

Therefore, we do not suggest that the primary dimension found in this report is representative of "primary" ADHD variation, nor that the focus of research should be only on this trait. Rather, the general ADHD dimension appears to be primarily a representation of ADHD items that relate to the impulsivity, hyperactivity, and susceptibility to immediate distractors, or more broadly stated, inhibition. The inattention factor specific dimension, on the other hand, represents the multistep processes with which children with ADHD struggle. The orthogonality of these two dimensions and the obvious differences in symptoms that load on each dimension likely forecast implications for treatment, understanding of executive functions in ADHD, and prediction of persistence of symptoms or functional outcomes.

ADHD Symptom Differences by Age and Gender

Fortunately, significant differential item functioning was not found across gender or age groups. This implies that there are, in fact, no systematic differences between age and gender groups regarding the probability of endorsement of specific ADHD symptoms given the same ADHD severity. However, mean and variance differences across gender and age clearly exist. Boys show greater severity on the hyperactivity specific dimension. As this factor was dominated by the item *runs about* and boys traditionally show more symptoms of hyperactivity/impulsivity (Biederman et al., 2002), this finding is consistent with the empirical literature. In addition, younger children were more severe on the general dimension and the hyperactivity specific dimension. This finding also supports existing research, as the general dimension here is primarily a representation of ADHD items that relate to inhibition or cognitive control. Evolving cognitive control with age is expected (Davidson, Amso, Anderson, &

Diamond, 2006) and would likely result in lower means of older children for the general and hyperactivity dimensions observed in this study.

Conclusion

Traditionally, studies have examined differences in ADHD as a function of subtype (i.e., inattentive vs. hyperactive/impulsive vs. combined type). The analyses presented in this report clearly show that subtype comparisons may be insufficient to explore differences in cognitive, academic, socio-emotional and functional domains. Few executive function (EF) differences have been found between children diagnosed with inattentive type and those diagnosed with combined type, and EF deficits specific to the hyperactivity/impulsivity subtype are rare (Willcutt, Doyle, Nigg, Faraone, & Pennington, 2005). Researchers have suggested the EF deficits and underachievement in academics is most closely tied to symptoms of inattention (Willcutt et al., 2012, 2005) and suggested that there is little evidence to support the validity of the hyperactivity/impulsivity domain (Willcutt et al., 2012). This report demonstrates that using the subtype model to examine the role of hyperactivity/impulsivity would be remiss. Clearly there are elements of inattention that are important and inextricably linked to hyperactivity and impulsivity.

Future Directions

The orthogonality of the inattention and hyperactivity and the obvious differences in symptoms that load on each dimension forecast implications for treatment, understanding of executive functions in ADHD, and prediction of persistence of symptoms or functional outcomes. Item factor scores can be generated from each subscale so that a clinician can simultaneously gauge an individual's severity on general ADHD trait variation (i.e. the primary ADHD dimension), in addition to their relative severity on the specific dimensions of inattention

and hyperactivity. Second, the bifactor model can help us to explore differences in etiological mechanisms and cognitive correlates for each cluster of symptoms. Further research can use the bifactor model to untangle the tenuous relationship between genetics, symptoms, and cognitive manifestations of the disorder such as planning, spatial working memory, and delay aversion (Castellanos, Sonuga-Barke, Milham, & Tannock, 2006). Finally, use of the restricted bifactor model for analysis of treatment effect may better inform treatment approaches that may be particularly effective for children with specific behavioral symptom presentation. Toward this goal, different scales designed to measure ADHD symptoms may be linked and therefore treatment findings directly compared among many clinical trials. The use of IRT to improve measurement, characterization, treatment, and implications of ADHD is just beginning. Future research using IRT has the opportunity to make substantial contributions to the field and positively impact the lives of many individuals diagnosed with ADHD.

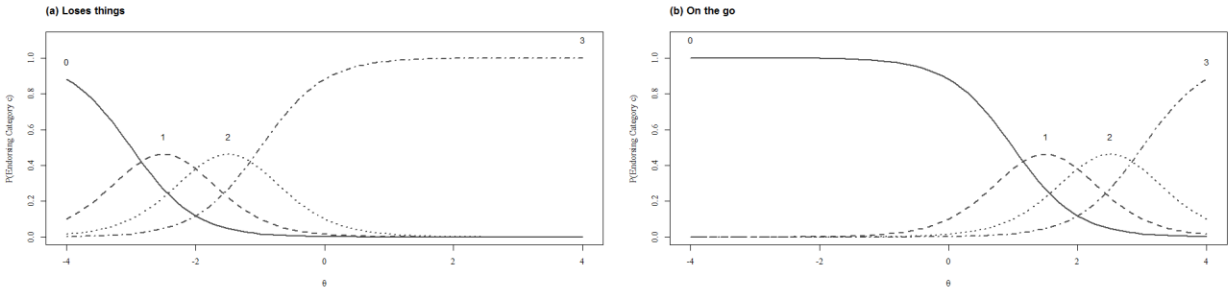

Figure 2-1. Graded model trace lines for two items with k = 4 response categories. Each line represents the corresponding probability of endorsing the kth category given θ (a unidimensional latent trait).

Table 2-1.

Demographic Characteristics for Development, Cross-validation, and Total Samples

	Development sample *n* = 1357		Cross-validation sample *n* = 255		Total sample *n* = 1612	
	n	*%*	*n*	*%*	*n*	*%*
Gender (% Male)	997	73.63	179	70.00	1176	71.46
Diagnoses						
ADHD	1329	98.15	249	97.65	1578	97.89
Inattentive	380	28.06	83	32.55	463	28.72
H/I	57	4.21	6	2.35	63	3.91
Combined	903	66.69	159	62.35	1062	65.88
DBDs (ODD&CD)	512	37.81	83	32.55	595	36.91
Anxiety disorders	353	26.07	31	12.16	384	23.82
OCD						
Affective Disorders	165	12.19	7	2.75	172	10.67
Tic Disorders						
PTSD						
Ethnicity						
Caucasian	1009	74.52	181	70.98	1190	73.82
Black or African-American	170	12.56	30	11.76	200	12.41
Asian	19	1.40	6	2.35	25	1.55
Other	42	3.10	15	5.88	57	3.54
Hispanic	115	8.49	23	9.02	138	8.56
	M (*SD*)	Range	*M* (*SD*)	Range	*M* (*SD*)	Range
Age	10.72 (2.57)	(5.98, 17.95)	10.42 (2.5)	(6.06, 17.11)	10.68 (2.56)	(5.98, 17.95)
ADHD-RS-IV-I	21.67 (4.66)	(0, 27)	21.2 (4.32)	(5, 27)	21.59 (4.61)	(0, 27)
ADHD-RS-IV-H/I	17.1 (6.69)	(0, 27)	15.78 (6.93)	(0, 27)	16.90 (6.75)	(0, 27)
ADHD-RS-IV-Total	38.78 (9.67)	(0, 54)	36.96(9.22)	(11, 54)	38.49 (9.62)	(0, 54)

Note. Other = mixed race, American Indian, or Alaska Native. ADHD-RS-IV-I = Inattentive, ADHD-RS-IV-H/I = Hyperactive/Impulsive.

Table 2-2.
Demographic Characteristics for Developmental Sample

	B4Z-MC-LYAC n = 364		B4Z-MC-LYAS n = 162		B4Z-MC-LYAQ n = 214		B4Z-US-LYBP n = 191		B4Z-US-LYCC n = 307		UCLA3 n = 37		UCLA4 n = 82	
	n	%	n	%	n	%	n	%	n	%	n	%	n	%
Gender (% Male)	266	73.08	144	88.89	154	98.72	126	65.97	222	72.31	22	59.46	63	76.83
Sites	14		14		20		14		14		1		1	
Diagnoses														
ADHD	348	95.87	155	95.68	211	98.6	189	98.95	307	100	37	100	82	100
Inattentive	118	33.43	57	35.85	53	24.8	47	24.61	66	21.5	10	27.03	29	35.37
H/I	7	1.98	5	3.14	5	2.34	2	1.05	7	2.28	27	72.97	4	4.88
Combined	228	64.59	97	61.01	153	71.5	142	74.35	234	76.22	0	0	49	59.76
DBDs (ODD&CD)	143	39.29	33	20.4	105	49.07	85	44.5	100	32.57	14	37.83	32	39.02
Anxiety disorders	11	3.02	17	10.5	96	44.86	190	99.48	4	.13	15	40.54	20	24.39
OCD	0	0	6	3.70			1	.52	0	0	0	0	0	0
Affective disorders	11	3.02	3	1.85	135	63.08	8	4.19	2	.65	0	0	6	7.32
Tic disorders			156	96.3							2	5.41	3	3.66
PTSD					9	4.21					0	0	0	0
Ethnicity														
Caucasian	259	71.15	144	88.89	181	85.38	155	81.15	210	68.4	18	48.64	42	51.22
Black or African-American	81	22.25	7	4.32	14	6.6	8	4.19	46	14.98	3	8.11	11	13.41
Asian	4	1.1	1	.62	1	.47			1	.33	5	13.51	7	8.54
Other	13	3.57	6	3.7	9	4.25	1	.52	8	2.61	3	8.11	2	2.44
Hispanic	7	1.92	4	2.47	7	3.3	27	14.14	42	13.68	8	21.62	20	24.39
	M (SD)	***range***	***M (SD)***	***range***	***M (SD)***	***range***	***M (SD)***	***range***	***M (SD)***	***range***	***M (SD)***	***range***	***M (SD)***	***range***
Age	11.26 (2.34)	(8.02, 17.85)	11.23 (2.39)	(7.00, 17.31)	11.38 (2.56)	(7.01, 17.81)	11.97 (2.69)	(8.01, 17.95)	8.89 (1.63)	(5.98, 11.97)	10.34 (1.33)	(8.13, 12.82)	9.82 (3.08)	(6, 17)
CGI-S	4.72 (.87)	(1, 7)	4.59 (.86)	(2, 7)	4.96 (.83)	(3, 7)	4.82 (.85)	(3, 7)	5.00 (.76)	(4, 7)	4.73 (.45)	(4, 5)	4.13 (1.14)	(1, 6)
ADHD-RS-inattentive	22.01 (4.42)	(1, 27)	21.40 (4.26)	(5, 27)	23.05 (3.49)	(12, 27)	22.01 (3.80)	(0, 27)	22.81 (3.51)	(6, 27)	18.46 (3.64)	(9, 26)	13.44 (6.14)	(0, 27)
ADHD-RS-hyperactive	16.88 (6.91)	(0, 27)	15.95 (7.02)	(0, 27)	17.79 (6.77)	(0, 27)	16.77 (6.23)	(11, 27)	19.54 (5.23)	(3, 27)	15 (4.85)	(5, 25)	11.17 (6.91)	(0, 24)
ADHD-RS-total	38.9 (9.69)	(2, 54)	37.35 (9.09)	(11, 54)	40.84 (8.55)	(18, 54)	38.82 (8.41)	(14, 54)	42.36 (6.84)	(23, 54)	33.46 (7.19)	(20, 46)	24.61 (11.90)	(0, 51)

Table 2-3.
Demographic Characteristics for Cross-validation Sample

	B4Z-MC-LYAT *n* = 196		UCLA1 *n* = 33		UCLA2 *n = 26*	
	n	*%*	*n*	*%*	*n*	*%*
Gender (% Male)	136	69.39	19	57.58	24	92.31
Number of Sites	9		1		1	
Current Diagnoses						
ADHD	190	96.94	33	100	26	100
Inattentive	77	39.29	3	9.09	3	11.54
H/I	3	1.53	2	6.06	1	3.85
Combined	110	56.12	28	84.85	21	80.77
DBDs (ODD&CD)	39	19.9	32	96.97	12	46.15
Anxiety disorders	8	4.08	20	60.61	3	11.54
OCD	0	0	0	0	0	0
Affective Disorders	7	3.57	0	0	0	0
Tic Disorders	NA	NA	0	0	0	0
PTSD	NA	NA	0	0	0	0
Ethnicity						
Caucasian	151	77.04	21	63.64	9	34.62
Black or African-American	26	13.27	1	3.03	3	11.54
Asian	2	1.02	1	3.03	3	11.54
Other (mixed race, american indian, alaska native)	10	5.1	5	15.15	0	0
Hispanic	7	3.57	5	15.15	11	42.31
	M (*SD*)	Range	*M* (*SD*)	Range	*M* (*SD*)	Range
Age	10.37 (2.53)	(6.06, 16.34)	10.44 (2.62)	(7.17, 17.11)	10.78 (2.07)	(7.01, 14.4)
CGI-S	4.61 (.72)	(2, 6)	4.78 (0.58)	(4, 6)	4.56 (.51)	(4, 5)
ADHD-RS-Inattentive	22.02 (4.08)	(5, 27)	17.94 (4.65)	(4, 25)	18.65 (4.14)	(9, 24)
ADHD-RS-Hyperactive	15.8 (7.38)	(0, 27)	15.62 (5.98)	(1, 23)	15.31 (5.07)	(2, 26)
ADHD-RS-Total	37.79 (9.57)	(11, 54)	33.56 (8.89)	(5, 45)	33.96 (7.45)	(19, 48)

Table 2-4.

Model Fit Statistics for Unidimensional, Correlated Factor, and Bifactor Models

Model	-2 log likelihood	AIC	BIC	M_2	df	RMSEA
Development						
1 factor	49744.23	49888.23	50263.57	4292.23	135	.15
2 factor	48077.12	48223.12	48603.67	857.05	134	.06
3 factor	47900.69	48050.69	48441.67	638.39	132	.05
Bifactor 2 specific dimensions	47702.21	47882.21	48351.39	449.28	117	.05
Bifactor 3 specific dimensions						
Bifactor restricted	47732.43	47902.43	48345.54	501.61	121	.05
Cross-validation						
Bifactor restricted	9424.74	9436.74	9457.99	386.89	201	.06

Note. 2 factor $r(SE)$ = .54(.02); 3 factor $r_{Iinatt\text{-}H}(SE)$: .53(.03), $r_{Inatt\text{-}Impuls}(SE)$: .49(.03), $r_{H\text{-}Impuls}(SE)$: .85(.03).

Table 2-5.

Model Fit Statistics for Competing Restricted Bifactor Models by Age and Gender

Restricted Bifactor Model	-2 Log Likelihood	AIC	BIC	M_2	df	RMSEA
Single group						
Impuls-Inatt	47732.43	47902.43	48345.54	2944.68	1346	.03
Inatt-Hyper	47751.38	47923.38	48371.7	2949.45	1345	.03
Age						
Impuls-Inatt	56968.33	57150.33	57640.33	4771.63	2771	.02
Inatt-Hyper	56976.27	57160.27	57655.65	4753.74	2770	.02
Gender						
Impuls-Inatt	57111.1	57293.1	57783.15	4856.26	2771	.02
Inatt-Hyper	57130.66	57314.66	57810.1	4858.22	2770	.02

Note. Inatt = Inattention, *Hyper* = Hyperactivity, *Impuls* = Impulsivity

Table 2-6.

Factor Loadings, Marginal Discrimination Parameters, and Thresholds for the Restricted Bifactor Model

Item	λ_1	λ_2	λ_3	a_1*	a_2*	a_3*	B_1	B_2	B_3
Careless mistakes	.4 (.05)	.62 (.04)	0	.74	1.34	0	-3.12	-1.67	-.21
Difficulty sustaining attention	.47 (.05)	.56 (.05)	0	.91	1.15	0	-3.19	-1.94	-.17
Does not listen	.51 (.05)	.39 (.05)	0	1.01	.72	0	-3.17	-1.49	.12
Fails to finish work	.35 (.05)	.7 (.04)	0	.64	1.67	0	-2.97	-1.76	-.31
Difficulty organizing	.31 (.06)	.68 (.04)	0	.55	1.58	0	-2.52	-1.51	-.35
Avoids task	.24 (.06)	.61 (.05)	0	.42	1.31	0	-2.88	-1.90	-.50
Loses things	.33 (.05)	.64 (.05)	0	.59	1.42	0	-1.98	-1.01	.10
Easily distracted	.55 (.05)	.57 (.05)	0	1.12	1.18	0	-3.46	-2.06	-.73
Forgetful	.31 (.05)	.65 (.04)	0	.55	1.45	0	-2.42	-1.22	.01
Fidgets	.66 (.04)	0	.23 (.07)	1.49	0	.40	-2.37	-1.45	-.33
Leaves seat	.64 (.04)	0	.38 (.06)	1.42	0	.70	-1.31	-.45	.51
Runs or climbs	.65 (.04)	0	.6 (.07)	1.45	0	1.28	-.72	-.02	.71
Difficulty playing quietly	.64 (.04)	0	.22 (.06)	1.42	0	.38	-1.05	-.11	.97
On the go	.74 (.03)	0	.36 (.06)	1.87	0	.66	-1.23	-.52	.25
Talks excessively	.72 (.04)	0	0	1.76	0	0	-1.55	-.71	.21
Blurts out answers	.74 (.03)	0	0	1.87	0	0	-1.38	-.45	.56
Difficulty awaiting turn	.79 (.03)	0	0	2.19	0	0	-1.29	-.49	.41
Interrupts	.83 (.02)	0	0	2.53	0	0	-1.75	-.92	.06

Table 2-7.

Discrimination and Intercept Parameters for the Restricted Bifactor Model

Item	a_1	a_2	a_3	c_1	c_2	c_3
Careless mistakes	1.01 (.10)	1.58 (.11)	0	5.86 (.34)	3.12 (.14)	.40 (.09)
Difficulty sustaining attention	1.17 (.10)	1.41 (.10)	0	5.84 (.34)	3.55 (.17)	.32 (.09)
Does not listen	1.12 (.08)	.86 (.08)	0	4.48 (.21)	2.11 (.1)	-.17 (.08)
Fails to finish work	.95 (.10)	1.90 (.13)	0	6.31 (.4)	3.73 (.19)	.66 (.10)
Difficulty organizing	.78 (.09)	1.72 (.12)	0	4.75 (.23)	2.85 (.14)	.66 (.10)
Avoids task	.53 (.08)	1.36 (.10)	0	4.21 (.19)	2.77 (.13)	.73 (.09)
Loses things	.81 (.09)	1.58 (.11)	0	3.52 (.16)	1.80 (.11)	-.18 (.09)
Easily distracted	1.53 (.13)	1.59 (.12)	0	7.63 (.55)	4.54 (.23)	1.62 (.12)
Forgetful	.75 (.09)	1.58 (.11)	0	4.24 (.2)	2.14 (.11)	-.02 (.09)
Fidgets	1.57 (.10)	0	.56 (.10)	3.95 (.21)	2.42 (.13)	.55 (.09)
Leaves seat	1.64 (.10)	0	.98 (.11)	2.51 (.13)	.86 (.09)	-.97 (.10)
Runs or climbs	2.36 (.24)	0	2.18 (.32)	2.32 (.24)	.07 (.13)	-2.29 (.24)
Difficulty playing quietly	1.49 (.09)	0	.52 (.08)	1.65 (.10)	.17 (.08)	-1.53 (.09)
On the go	2.21 (.13)	0	1.07 (.12)	3.01 (.16)	1.28 (.12)	-.62 (.11)
Talks excessively	1.74 (.10)	0	0	2.7 (.13)	1.24 (.09)	-.37 (.09)
Blurts out answers	1.85 (.11)	0	0	2.55 (.12)	.84 (.09)	-1.03 (.10)
Difficulty awaiting turn	2.20 (.13)	0	0	2.84 (.14)	1.07 (.11)	-.91 (.1)
Interrupts	2.58 (.15)	0	0	4.52 (.22)	2.37 (.15)	-.16 (.11)

3. CHAPTER 2: The use of item response theory for the evaluation of medication effects in clinical trials of pediatric ADHD: Increasing precision of measurement.

3.1. Introduction

An ADHD diagnosis confers significant risk for future psychopathology (e.g. anxiety and depression; Lahey et al., 2007), substance abuse (Molina et al., 2013), poor self-esteem (Harpin, Mazzone, Raynaud, Kahle, & Hodgkins, 2016; Shaw et al., 2012), and poor social functioning (Harpin et al., 2016; Shaw et al., 2012). Intervention can help to mitigate many negative long-term outcomes, but individuals often struggle to find appropriate treatment (Harpin et al., 2016; Shaw et al., 2012). Therefore, it is essential that effective and targeted treatment approaches for children with ADHD continue to be evaluated and pursued.

Improving prediction of treatment response is receiving increased attention in ADHD research, and mental health more generally. Individuals seeking treatment for any mental health condition may try several treatment approaches before finding a pharmacological, behavioral, or combination treatment that maximizes efficacy while minimizing adverse effects. Toward this goal, researchers aim to develop a more "personalized" approach to medicine through targeted intervention (Insel, 2015). In other words, which treatments work and for whom? Identification of targeted treatment approaches also requires analytic methods that appropriately identify the effect of treatment, while accounting for the noise of measurement error. The present study aims to improve the way we estimate treatment effect for children and adolescents with ADHD by comparing the use of an advanced statistical method (item response theory) to traditional methods of modeling treatment (ANOVA).

Pharmacotherapy

Recent estimates suggest that approximately 32% of children diagnosed with ADHD have received consistent treatment in the past 12 months (Froehlich et al., 2007), with varying

estimates depending on location of the survey within the U.S. and definition of treatment period (Angold, Erkanli, Egger, & Costello, 2000; Jensen et al., 1999; Wolraich, Hannah, Baumgaertel, & Feurer, 1998). Practice parameters recommend that the first line pharmacological treatment for ADHD be approved by the Food and Drug Administration (Pliszka, 2007). Dextroamphetamine (DEX), D- and D,L-methylphenidate (MPH), mixed salts amphetamine, atomoxetine, and most recently non-stimulants guanfacine and clonidine are current FDA approved drugs for the treatment of ADHD.

Stimulant pharmacotherapy.

Stimulant medications, including DEX, MPH, and mixed salts amphetamine, increase the release and/or inhibit reuptake of dopamine and norepinephrine, two neurotransmitters essential for attention and cognition (Arnsten, Scahill, & Findling, 2007). An estimated 65% to 75% of children and adolescents treated with stimulants are responders relative to 4% to 30% who respond to placebo (Pliszka, 2007). The effect of stimulant treatment is also profound, evidenced by effect sizes averaging around 1.0 (Pliszka, 2007). The two stimulant types, MPH and amphetamines, have been shown thus far to produce very similar rates of response in individuals with ADHD. Approximately 41% of individuals respond equally to MPH and amphetamines, while 44% respond preferentially to one or the other (Arnold, 2000).

Non-stimulant pharmacotherapy.

Inadequate response to stimulants coupled with severe side effects for some individuals necessitated the development and use of non-stimulant pharmacotherapy for the treatment of ADHD (Childress & Sallee, 2014). Clonidine, guanfacine, and atomoxetine are three non-stimulants that are increasingly used as monotherapies or adjunct therapy to psychostimulants (Childress, 2012). Atomoxetine (Cheng, Chen, Ko, & Ng, 2007; Michelson et al., 2002) is a

selective norepinephrine-reuptake inhibitor, while guanfacine (Biederman et al., 2008; Sallee, Lyne, Wigal, & McGough, 2009) and clonidine (Jain, Segal, Kollins, & Khayrallah, 2011) are selective α_2-adrenergic agonists.

Clinical trials of a number of pharmacotherapies including lisdexamfetamine (Biederman, Krishnan, Zhang, McGough, & Findling, 2007), OROS and IR methylphenidate (Wolraich et al., 2001), atomoxetine (Michelson et al., 2001), and guanfacine (Scahill et al., 2001) have reported significant and proportionally uniform reductions in symptoms of hyperactivity/impulsivity and inattention. Effect sizes are typically used to compare results between studies that use a number of different measurement scales. Non-stimulants are generally less effective than stimulants (Wigal, 2012). Effect sizes range from around a Cohen's d of 0.6 for non-stimulants, to appromixately 1.0 for stimulants (Faraone, Biederman, Spencer, & Aleardi, 2006; Faraone & Buitelaar, 2009; Hirota, Schwartz, & Correll, 2014). Although treatment effect is robust in ADHD, there remains substantial variability in individual response to treatment. Approximately 20% to 30% of individuals do not respond to a first-line stimulant treatment and estimates vary by pharmacotherapy (Childress & Sallee, 2014). Clearly, the various pharmacotherapies available for ADHD are not equally effective for every child. In addition, we have yet to identify clear predictors of treatment response to minimize the guess and check approach frequently used through the course of intervention.

Therapeutic Effects on ADHD Symptom Dimensions

Preferential response to specific drug formulations (Arnold, 2000) and the effective use of non-stimulants as adjunctive therapies (Hirota et al., 2014) suggest that different medications may differentially impact the symptoms or subtypes of ADHD. However, meta-analyses of a number of stimulant and non-stimulant compounds have found comparable decreases in

symptoms of inattention and hyperactivity for each treatment (Faraone & Buitelaar, 2009; Hirota et al., 2014). Variability of drug effectiveness coupled with overall uniform effect sizes across symptom subtypes are contradictory and it is therefore apparent that a piece of the puzzle is missing. The way that ADHD symptoms are currently modeled may affect our ability to tease apart the unique effect of varied treatment approaches. Further exploration of the comparative effect of stimulant and non-stimulant medications will significantly advance our understanding of treatment of ADHD.

Important predictors of treatment response have been established in some domains (e.g. genetics; McGough et al., 2006), but researchers have struggled to consistently find unique moderators of treatment. Inconsistency in our ability to predict response to treatment suggests that the way we define ADHD and model symptom improvement may contribute to the lack of findings. There are a number of statistical strategies used to model symptom change over time and treatment effect to varying success. Item response theory (IRT) is one statistical method that may improve the precision with which we estimate the effect of treatment. IRT allows the simultaneous modeling of the scale of measurement in addition to change in the underlying latent trait (McArdle, Petway, & Hishinuma, 2015). Previous studies have also demonstrated that traditional regression models can underestimate estimates of effect size relative to IRT models (Cai, Choi, Hansen, & Harrell, 2016).

Using IRT to Model Treatment Effect

The potential benefits of modeling treatment effects in the IRT framework extends beyond increased precision of estimates by allowing for the multidimensionality of a construct. Currently, clinical trials of ADHD pharmacotherapy model differences in treatment response by creating summed scores from the items of a scale. However, researchers have consistently found

the unidimensional model of ADHD symptoms to be a poor representation of the construct (Martel, von Eye, & Nigg, 2010; Toplak et al., 2009). The bifactor model has thus far been the superior performer when comparing the model fit of ADHD symptoms (Toplak et al., 2012; Ullebø et al., 2012). The bifactor model allows extraction of general trait variability, leaving precise specific dimensions that model remaining dependence between items (see Figure 3-1 for diagram). Rather than general ADHD trait variation, the general ADHD dimension appears to be primarily a representation of ADHD items that relate to cognitive control and inhibition; representative of the shared variability between symptoms of hyperactivity, impulsivity and inattention (see Study 1). What remains are two specific dimensions orthogonal to (or unique from) the primary dimension; one including inattention symptoms and the other symptoms of hyperactivity. The flexibility of the bifactor model is clearly beneficial in the modeling of ADHD symptoms, where in fact impulsivity items only load on the primary dimension (see Study 1). Parsing ADHD symptom variance into a general dimension and specific dimensions allows researchers to examine which dimensions of the ADHD construct are important predictors of psychosocial, academic, and cognitive functioning, in addition to treatment response.

Item response theory (IRT) also offers a novel means by which longitudinal data and treatment effect can be measured. Using IRT, the bifactor model can be used to estimate change in all orthogonal dimensions from baseline to treatment endpoint. This type of analysis may yield a more precise understanding of the effect of different treatments on the orthogonal dimensions of ADHD. For example, one treatment (treatment A) may provide greater symptom reduction in the inattention domain, while another treatment (treatment B), may have a greater effect on the hyperactivity subdomain. Different patterns of change between the general dimension and

specific subdimensions may help to inform our understanding of the mechanism of various medications.

Use of more effective statistical models, such as IRT, in treatment research was previously limited by sample size. Estimation of change in an IRT framework requires estimation of many item parameters. Fortunately, the advent of data sharing in psychiatry is well-timed and can help to solve problems of sample size. Calibrated item parameters can be generated from a large pre-treatment sample and then used in subsequent analyses such that fewer parameters need be estimated. The use of more advanced statistical methodology coupled with large sample sizes available through data sharing can help us to answer these new and interesting questions that will help to advance our understanding of ADHD treatment.

The central aims of the present study are to (1) use a multiple group longitudinal IRT model to estimate change in the general, inattention, and hyperactivity dimensions from baseline to treatment endpoint for two different treatment trials (non-stimulant versus placebo and additive benefit of a non-stimulant to the industry standard stimulant treatment), (2) estimate ANOVA models of change in symptom summed scores for reference and (3) demonstrate that IRT can be useful in estimating change in psychological constructs and identifying differences in treatment sensitivity as defined by bifactor dimension mean differences.

3.2. Methods

Study 1: Guanfacine and DMPH

Participants.

Participants were sampled from the Translational Research to Enhance Cognitive Control (TRECC) treatment study that consisted of four sub-projects with a total of 502 participants. Participants and their families were recruited from the greater Los Angeles area through flyers,

radio ads, and school and health practitioner referrals. One of these sub-projects, Project 1, examined the impact of treatment on cognitive control in children and adolescents with ADHD (McCracken et al., under review). Data from 139 participants from Project 1 will be used for the purpose of these analyses. To be eligible for participation in Project 1, children had to be ages 7 to 14, meet DSM-IV criteria for any one ADHD subtype (inattentive, hyperactive/impulsive, or combined), and have a full scale IQ greater than 70. Children with a diagnosis of an autism spectrum disorder, major depressive disorder (MDD), psychosis, bipolar disorder, or chronic tic disorder were excluded from participation. The reduced sample was 66% (*n* = 92) Caucasian and 69.8% male (*n* = 97).

Procedure.

The TRECC study was an 8-week randomized, comparative parallel-group study with three active treatments: (1) guanfacine, (2) psychostimulant (d-methylphenidate extended-release [DMPH]), or (3) combined (COMB) treatment. Psychostimulants (e.g. DMPH) continue to be considered the gold standard treatment due to their superior efficacy. However, a new focus has emerged on identifying whether combination treatments, such as guanfacine added to a stimulant, may result in enhanced treatment outcomes. For this reason, only the DMPH and combined treatment groups were included in these illustrative analyses. At baseline, individuals were randomized to one of the three treatment groups. Treatments were applied sequentially: the first 4 weeks, subjects received guanfacine or placebo, and then at Week 5, subjects continued receiving guanfacine alone or added a stimulant (DMPH). Guanfacine was initiated at 0.5mg BID and increased by 0.5mg increments to a total possible dose of 1.5mg BID. In the DMPH group, those < 25 kg started at a dose of 5mg once daily and increased in 5 mg increments to a

total daily dose of 20 mg. The ADHD-RS-IV was administered at baseline and week 8 treatment endpoint.

Study 2: Atomoxetine

Two identical studies of atomoxetine were also examined (HFBK *n* = 144; HFBD *n* = 147; Spencer et al., 2002). Both were 9-week, multicenter, randomized, stratified, parallel, double-blind studies of tomoxetine hydrochloride, methylphenidate hydrochloride, and placebo in children ages 7 to 12 diagnosed with ADHD. Children were first stratified into two groups; (1) stimulant naïve and (2) prior stimulant exposure. Children who were stimulant naïve were randomized to either tomoxetine hydrochloride, methylphenidate hydrochloride, or placebo (3:3:2 randomization ratio) while those who had prior stimulant exposure were randomized to tomoxetine hydrochloride or placebo (1:1 randomization ratio). Dose ranged from 5mg to 45mg BID for tomoxetine hydrochloride (max 2mg/kg/day) and 5mg to 30 mg for methylphenidate hydrochloride (max 1.5/kg/day). Both medications were increased in 5mg increments each week over the course of 9 weeks until a maximum tolerated daily dose was established. All studies had a flexible dosing schedule. For the purpose of these analyses, the HFBD and HFBK samples were combined as was done in the original paper (Spencer et al., 2002).

To be eligible for randomization, participants had to meet DSM-IV diagnostic criteria for ADHD as assessed by the K-SADS-PL (Kaufman et al., 1997), have an ADHD-RS-IV score 1.5 standard deviations above gender and age norms for either the inattention or hyperactivity/impulsivity subscale or total for combined type. Patients were excluded if they were poor metabolizers of CYP2D6, weighed less than 25 kg, were currently taking a psychotropic medication, had a history of drug or alcohol abuse in the last three months, a history of bipolar disorder, psychosis, any organic brain disease, seizure disorder or any

significant prior medical condition. All parents or legal guardians of patients provided written informed consent prior to participation. All study procedures were approved by each site's investigational review board.

Measures

ADHD-RS-IV.

The ADHD Rating Scale–IV (ADHD-RS-IV; DuPaul et al., 1998), as described in Study 1, is a clinician-rated measure based on the ADHD diagnostic criteria from the DSM-IV-TR (American Psychiatric Association, 2000). The ADHD-RS-IV contains nine inattention and nine hyperactivity-impulsivity items that are presented in alternating order. Of the nine hyperactivity/impulsivity items, five items reflect hyperactivity symptoms and four reflect impulsivity symptoms. In each of the samples, a parent or legal guardian was asked by a trained clinician to rate the frequency of their child's behavior at baseline and at treatment endpoint. Each item was rated on a 4-point Likert scale, where 0 = *never or rarely*, 1 = *sometimes*, 2 = *often*, 3 = *very often*, resulting in a minimum score of 0 and a maximum score of 27 for each subscale. At the time of the baseline assessment, children and adolescents were unmedicated and at treatment follow-up, had reached an optimally titrated dose of study medication. For the present investigation, summed scores were computed for the nine ADHD-RS-IV inattention items (ADHD-RS-IV-IN), the nine hyperactivity/impulsivity items (ADHD-RS-IV-H/I), the five hyperactivity items (ADHD-RS-IV-H) and a total score based on all 18 items (ADHD-RS-IV-T).

Statistical Analyses

Descriptives.

Descriptive information including age, gender and diagnostic information was computed for each treatment group in both studies (Table 3-1).

ANOVA.

Four 2 x 2 mixed factorial designs were computed for each study (a total of eight analyses) to compare the effect of atomoxetine versus placebo and DMPH versus combined treatment (DMPH + guanfacine) on ADHD-RS-IV-IN, ADHD-RS-IV-H/I, ADHD-RS-IV-H, and ADHD-RS-IV-T summed scores using SPSS version 22. Terms included in the model were time (baseline and treatment endpoint), treatment group, and a time by treatment interaction. As the purpose of these analyses was to compare change over time between treatment groups, only time by treatment interactions are reported in Table 3-3 for atomoxetine versus placebo and Table 3-4 for DMPH versus COMBO. All statistical tests were performed using a two-sided test at an alpha level of .05. Corrected tests for unequal variances were used when appropriate based on Levene's test for equality of variances. Effect sizes were calculated using Cohen's *d*.

IRT.

The sample sizes of each treatment study were not sufficiently large to both estimate parameters *and* change over the three latent variables of interest (general, inattention, and hyperactivity dimensions). To address this issue, item discrimination and threshold parameters from a prior study using the ADHD-RS-IV with the same single time point IRT model structure were used (see Study 1). This reduced the number of parameters estimated using the smaller treatment samples and also serve to fix the scale of the dimensions so that both baseline and treatment endpoint means and variances for the latent dimensions could be estimated.

Models were estimated in flexMIRT version 3 using the Metropolis-Hastings Robbins-Monro (MH-RM) algorithm (Cai, 2010a, 2010b) and were verified using the Bock-Aitkin EM algorithm (Bock & Aitkin, 1981). Graded response models (GRM; Samejima, 1969) with a normally distributed latent trait were used to model the ordinal nature of the ADHD-RS-IV four

category responses. A benefit of IRT is the use of full information maximum likelihood for estimation of parameters, which allows for the inclusion of individuals who have missing item responses in estimation. However, because the purpose of these analyses was to compare interpretation of traditional ANOVA results to results obtained from an IRT analysis, only complete cases were used due to listwise deletion in ANOVA.

Two multiple-group, longitudinal IRT models were estimated, one for Study 1 and one for Study 2. Each treatment arm was classified as one group, such that each longitudinal IRT model contained two groups (i.e. placebo and atomoxetine in study 1; DMPH and DMPH+guanfacine in study 2). In each model, two restricted bifactor models were included to represent the multidimensional structure of ADHD items at baseline and treatment endpoint. Therefore, one bifactor model included all baseline ADHD items and the second included all treatment endpoint ADHD items. In the restricted bifactor model (as described in Study 1), all 18 items loaded on one general dimension. In addition, the nine inattention items (e.g. *loses things*) also loaded on one specific factor and the five hyperactivity items (e.g. *driven by a motor*) loaded on a second specific factor. The four impulsivity items (e.g. *interrupts or intrudes*) loaded exclusively on the general dimension (see Figure 3-1). Item discrimination and threshold parameters were all fixed to values obtained from the restricted bifactor model of the ADHD-RS-IV items fit in Study 1. All six factor means and variances (two of each general, inattention, and hyperactivity) were freely estimated, in addition to three covariance parameters between the baseline and treatment endpoint general dimensions, inattention dimensions, and hyperactivity dimensions, respectively. The remaining covariances were fixed to zero. Significance of mean difference between treatment dimension means was calculated using the Wald test, producing a *z* score. To gauge significance of mean differences between baseline and treatment endpoint for

each dimension, Cohen's *d* effect sizes were computed between treatment groups within Study 1 and Study 2. IRT scaled scores (Expected *A Posteriori* (EAP) scores) were computed for each of the baseline and treatment endpoint dimensions as posterior means (Thissen & Wainer, 2001).

Differences between results from the mixed factorial and longitudinal IRT models were evaluated using Cohen's *d* effect sizes. It is important to note that results from the mixed factorial design and longitudinal IRT models are not directly comparable. Means generated from the mixed factorial and IRT models represent different constructs as the purpose of the IRT model is to parse out *unique* sources of variability in items while the summed score method does not. The distinction between the models informed interpretation of results.

3.3. Results

Study 1: Atomoxetine

Participants.

A total of 231 participants completed the baseline assessment, 129 in the atomoxetine group and 124 in the placebo group (see Table 3-1). For more information regarding participants, see Spencer et al., 2002. Table 3-3 includes raw summed baseline and treatment endpoint ADHD-RS-IV scores for inattention (ADHD-RS-IV-I), hyperactivity (ADHD-RS-IV-H), hyperactivity/impulsivity (ADHD-RS-IV-H/I), and total symptoms (ADHD-RS-IV-T) for participants with complete observations.

A total of 195 participants completed the treatment endpoint assessment, 102 in atomoxetine and 93 in placebo. For these analyses, only participants who completed both baseline and treatment endpoint assessments were included (atomoxetine: $n = 102$, placebo: $n = 93$). A total of 27 participants in the atomoxetine group and 31 participants in the placebo group were not included due to missing observations at treatment endpoint.

ANOVA.

A significant effect of time by treatment for ADHD-RS-IV-I ($F(1, 192) = 27.83$, $p < .001$), ADHD-RS-IV-H ($F(1, 193) = 32.50$, $p < .001$), ADHD-RS-IV-H/I ($F(1, 193) = 35.97$, $p < .001$) and ADHD-RS-IV-T ($F(1, 192) = 36.93$, $p < .001$) was observed. Examining mean differences, the atomoxetine group experienced a greater decrease in all ADHD-RS-IV summed scores relative to the placebo group. Cohen's *d* estimates of effect size measuring the difference between mean improvement of total, inattention, and hyperactivity/impulsivity symptoms ranged from .77 to .87, representative of a large effect.

IRT.

The longitudinal IRT analyses revealed significant decreases in means for the general dimension ($z = 6.12$, $p < .001$) and the inattention dimension ($z = 3.64$, $p < .001$), but not the hyperactivity dimension ($z = 1.15$, $p = .25$). The general dimension was most effectively treated by atomoxetine ($d = .66$) with a medium-large effect size, followed by the inattention dimension ($d = .51$) with a medium effect size and finally the hyperactivity dimension ($d = .15$) with a small effect size.

In the atomoxetine group, there was differential improvement in latent variable means across the three dimensions such that greatest improvement was seen in the general dimension, followed by the inattention dimension and the hyperactivity dimension. The mean of the general dimension decreased by 1.23 standard deviations, the inattention dimension decreased by 1.06 and the hyperactivity dimension decreased by .37 standard deviations. The placebo group, in contrast, experienced fairly uniform symptom reduction, ranging between .23 and .42 standard deviations on the three domains. Estimates of effect size for comparison of mean differences

between the atomoxetine and placebo group reflect the general pattern of reduced dimension means.

IRT: Consistency of treatment effect.

Comparing the correlations between dimensions across time points for different treatment groups can also provide information about treatment effect. If correlations are conceptualized as maintenance of rank order of subjects between baseline and treatment endpoint, high correlations indicate that the order of subjects has been maintained while low correlations indicate re-ordering of subjects between treatment endpoint relative to baseline. Greater re-ordering of subjects at treatment endpoint suggests that the treatment is not equally effective across all subjects. Table 3-2 includes variances (diagonal elements) and correlations (off-diagonal elements) between each dimension. While the means of the inattention dimensions decreased more substantially for the atomoxetine group, the correlations between baseline and treatment endpoint for the placebo and atomoxetine group are nearly identical ($r_{atomoxetine}$ = .48; $r_{placebo}$ = .45). This may suggest a similar degree of subject re-ordering between atomoxetine and placebo from baseline to treatment endpoint. In contrast, the correlation between baseline and treatment endpoint for the general dimension was stronger for placebo than for atomoxetine ($r_{atomoxetine}$ = .62; $r_{placebo}$ = .78). The substantially smaller variance for the atomoxetine treatment endpoint general dimension latent variable (s = 1.42, SE = .21) compared to the placebo group (s = 2.46, SE = .39) suggests that there is considerable variability in treatment endpoint scores for the general dimension across children. Finally, the correlations between baseline and treatment endpoint for the hyperactivity dimension were fairly consistent across the atomoxetine and placebo groups ($r_{atomoxetine}$ = .71; $r_{placebo}$ = .81). Therefore, treatment with atomoxetine did not

have a substantial impact on the hyperactivity dimension, and the rank order of subjects was maintained on the general dimension between baseline and treatment endpoint.

Study 2: TRECC

Participants

A total of 137 participants completed the baseline assessment, 69 in the DMPH group and 68 in the COMBO group (see Table 3-1). For more information regarding participants, see McCracken et al. (under review). Table 3-4 includes raw summed baseline and treatment endpoint ADHD-RS-IV scores for inattention (ADHD-RS-IV-I), hyperactivity (ADHD-RS-IV-H), hyperactivity/impulsivity (ADHD-RS-IV-H/I), and total symptoms (ADHD-RS-IV-T) for participants with complete observations. At treatment endpoint, a total of 116 participants completed the endpoint assessment, 60 in DMPH and 57 in COMBO. One participant in the COMBO group completed treatment endpoint, but not the baseline assessment. For these analyses, only participants who completed both baseline and treatment endpoint assessments were included (DMPH: $n = 60$, COMBO: $n = 56$).

ANOVA.

No significant effects of time by treatment for ADHD-RS-IV-I ($F(1, 114) = 1.38$, $p = .24$), ADHD-RS-IV-H ($F(1, 114) = 1.58$, $p = .21$), ADHD-RS-IV-H/I ($F(1, 114) = 2.81$, $p = .10$) and ADHD-RS-IV-T ($F(1, 114) = 2.57$, $p = .11$) were observed. Examining mean differences, the atomoxetine group experienced a greater decrease in all ADHD-RS-IV summed scores relative to the placebo group. Cohen's *d* estimates of effect size measuring the mean difference between total, inattention, and hyperactivity/impulsivity symptoms for COMBO compared to DMPH ranged from .22 to .33, representative of a small effect.

IRT.

The longitudinal IRT analyses revealed a significant decrease in the dimension mean for the general dimension ($z = 2.02$, $p = .04$), but no significant mean differences for the inattention dimension ($z = 1.08$, $p = .28$), or the hyperactivity dimension ($z = 0.67$, $p = .50$). Examining mean differences, COMBO was more effective than DMPH in treating the general dimension ($d = .39$) evidenced by a small-medium effect size, followed by the inattention dimension ($d = .25$) with a small effect size, and finally the hyperactivity dimension ($d = .14$) with a small effect size.

In both the COMBO and DMPH groups, there was differential improvement in latent variable means across the three dimensions such that greatest improvement was seen in the inattention dimension, followed by the general dimension and the hyperactivity dimension. In the COMBO group, the mean of the inattention dimension decreased by 1.49 standard deviations, the general dimension decreased by 1.09 standard deviations, and the hyperactivity dimension decreased by .67 standard deviations. The DMPH group experienced a similar pattern of mean decrease in latent variable means. The mean of the inattention dimension decreased by 1.27 standard deviations, the general dimension by .77 standard deviations, and the hyperactivity dimension by .49 standard deviations.

IRT: Consistency of treatment effect.

The patterns of correlations between baseline and treatment endpoint and variances also differed for COMBO and DMPH. Although COMBO produced a significantly larger mean decrease than DMPH on the general dimension, DMPH was more uniform in its treatment of the general dimension. The rank order of subjects between baseline and treatment endpoint was more uniform for DMPH ($r_{DMPH} = .73$; $r_{COMBO} = .58$) and variability in scores was much lower at treatment endpoint for DMPH ($\sigma_{DMPH} = .31$, $SE = .06$) than for COMBO ($\sigma_{COMBO} = 1.02$, $SE =$

.21). In contrast, for the inattention dimension, rank order of subjects between baseline and treatment endpoint was more uniform for COMBO treatment (r_{COMBO} = .61) than DMPH (r_{DMPH} = .31), however there was a similar degree of variability at treatment endpoint (σ_{DMPH} = .75, *SE* = .15; σ_{COMBO} = .82, *SE* = .17). This suggests that although there were no significant mean differences on the general dimension, COMBO treatment produced more uniform decreases in scores compared to the DMPH group. Finally, there were no mean differences in treatment effect between DMPH and COMBO for the hyperactivity dimension and correlations were relatively similar (r_{DMPH} = .68; r_{COMBO} = .51), suggesting consistency in overall effect and variability in effectiveness across subjects. It should be noted that although baseline variances look strikingly different between DMPH and COMBO for the general and inattention dimensions, these differences are insignificant based on 95% confidence intervals of the estimates (*general*: DMPH [0.43, 0.97], COMBO [0.67, 1.57]; *inattention*: DMPH [0.54, 1.24], COMBO [0.86, 2.00]).

Comparing Atomoxetine and DMPH versus COMBO: Summary

Another benefit of modeling treatment in IRT is the ability to make general comparisons in mean differences in dimension change across studies. While we cannot determine the significance of mean differences between studies as they were analyzed separately, general comparisons can be made. Because item parameters are fixed at the same values, the underlying latent traits of general ADHD, inattention, and hyperactivity are placed on the same scale. While effect sizes tell us that largest between-treatment differences are consistently seen on the general ADHD dimension, overall reduction in latent variable means reveal a differential pattern of improvement. In the atomoxetine trial, the general dimension showed the greatest decrease in the mean latent trait. In contrast, DMPH and COMBO showed the most impact on the inattention latent dimension.

In addition, the general pattern of changes in variability from baseline to treatment endpoint differs between the two studies. While between subject variability in scores decreases for the DMPH and COMBO groups, variability increases for the atomoxetine and placebo groups (see Figure 3-2). Increasing variability in scores at treatment endpoint may be indicative of greater variability in the effectiveness of treatment.

3.4. Discussion

The primary aims of this study were to use the modeling precision of IRT to examine the effect of ADHD medication treatment in studies of atomoxetine compared to placebo and DMPH compared to DMPH with guanfacine as an adjunctive treatment and to make general comparisons to results obtained from a classic ANOVA. This study revealed that estimating change in ADHD symptoms in the context of a multidimensional IRT model can provide significantly more information about treatment effect and the variability in effectiveness compared to ANOVA. The restricted bifactor model used in these analyses consists of one primary dimension representing impulsivity, hyperactivity, and susceptibility to immediate distractors (inhibition) and two specific dimensions, aspects of inattention and hyperactivity unexplained by the general dimension. Changes in latent variable means coupled with correlation and variance information reveal that symptoms are in fact not targeted equally when orthogonal dimensions are extracted from the ADHD symptoms. Differential improvement in mean scores across the three dimensions provides support not only to the proposed restricted bifactor structure, but also for the use of IRT in the analysis of treatment effect.

Multidimensionality of ADHD and Estimates of Treatment Effect

Researchers have established that ADHD symptoms are not best explained by a unidimensional construct (Toplak et al., 2009). In the present study, it is clear that ANOVA can

provide information about overall reduction in symptoms, but inherent flaws exist in using ANOVA to best decompose the differential effect of treatment on types of ADHD symptoms. In fact, the uniform reductions that have been observed in treatment studies with both stimulant and non-stimulant medications (Hirota et al., 2014) were obscured due to insufficient modeling of the multidimensional and shared variability among ADHD symptoms. In the analysis of summed scores, subdomains (e.g. inattention and hyperactivity/impulsivity) are not conditionally independent. This means that to some extent, the two subscales capture some of the same variability in symptoms. The bifactor model allows us to examine change in conditionally independent dimensions. Creating conditionally independent dimensions clearly allowed for more precise modeling of treatment effect. In other words, how do different treatments differentially affect the shared underlying features of ADHD symptoms?

Treatment Findings from IRT Analysis

Atomoxetine.

Atomoxetine produces significant reductions in inattention and the general dimension, however there is more variability in children's response to atomoxetine as measured by the general dimension. Atomoxetine therefore may provide an effective treatment for targeting inattention in children, but less consistently targets inhibition (hyperactivity, impulsivity, and susceptibility to immediate distractors). In addition, uniform reductions across all three dimensions were found for the placebo group. The ability to detect uniform reductions in symptoms due to placebo effect while teasing apart differential effectiveness of a medication across the three symptom dimension reveals the utility of IRT in examination of treatment effect.

DMPH and Guanfacine combination treatment.

While the ANOVA model revealed no significant effect of combination treatment over DMPH, the IRT model revealed that combination treatment is superior in the treatment of the general dimension. Examining variability in average scores provided by the IRT analysis, it appears that combination treatment may provide an effective treatment for targeting inattention symptoms in children, while DMPH may provide an effective treatment for targeting symptoms consistent with the general dimension (inhibition). These findings suggest that the recommended treatment approach may differ depending on a child's symptom presentation.

It is important to note that the results from ANOVA and the latent variable are not directly comparable. Change in inattention captured by a summed score is not the same as change captured by the inattention latent variable. Nonetheless, ANOVA results were provided for illustrative purposes in order to give a frame of reference for interpretation of scores.

Treatment of hyperactivity dimension.

Results from these two treatment studies also indicate that, while hyperactivity on the general factor is effectively treated by pharmacotherapeutic approaches, the hyperactivity subdimension is not well-treated. Generally, DMPH and combination treatment (DMPH + guanfacine) were more effective treatments for the hyperactivity subdimension, but symptom reduction was still low. When considering target behaviors for treatment, it is difficult to conceptualize how the behaviors manifested by the hyperactivity dimension (primarily *runs about or climbs*) may be distinct from the general dimension. However, further exploration of behavioral, cognitive, and biological correlates may help to highlight the uniqueness of this dimension. This would then inform our ability to operationalize change and better treat children with specific elevations in this symptom dimension.

Finally, IRT also provided information about general differences in how different drugs uniquely target the specific dimensions of ADHD. In the atomoxetine trial, the general dimension (representative of inhibition) was most effectively treated, followed by the inattention dimension and then the hyperactivity dimension. In contrast, DMPH and DMPH with guanfacine as an adjunctive treatment most effectively treated inattention, followed by the general dimension and finally hyperactivity. While definitive conclusions cannot be drawn across the atomoxetine and DMPH studies as the two were not included in the same analysis, future research that includes all treatments in one model can make comparative conclusions regarding the observed effects.

Future Directions

Despite substantial research exploring medication and behavioral intervention approaches for children with ADHD, little progress has been made in identifying moderators and/or mediators of treatment response. Genetics (Pliszka, 2012), comorbid psychopathology such as anxiety, and ADHD severity (Hinshaw, 2007), among others, have all been identified as treatment moderators, but results are inconclusive and highly inconsistent. The present investigation reveals that some of the apparent difficulty in identifying predictors of treatment response may have been a result of the way that the relationship between items was represented. As a next step, covariates can be included in the longitudinal IRT model to try and predict who might respond best to specific treatment approaches.

Investigators have also explored the effect of pharmacotherapy on executive functions in children with ADHD. Thus far, studies have found minimal differences in EF improvement between ADHD subtypes after treatment with medication (Solanto et al., 2009). Given the present findings, perhaps current operationalizations of subtypes impede our ability to tease apart

cognitive patterns of response to various pharmacotherapies. Future research can use the modified structure of ADHD symptoms to compare and predict changes in EF as a function of treatment.

Conclusion

The field of psychiatry, and more broadly treatment research, could benefit substantially from continued use of IRT models. IRT may contribute both to greater refinement of the measures that are used for treatment outcome; in addition to the evaluation of treatment effect in a model that better explains sources of variance and covariance among variables. For IRT models of ADHD symptoms to maximally inform our ability to treat children and adolescents, it is essential that cognitive, genetic and behavioral correlates of the specific dimensions be established. Creating behavioral, neurocognitive, or genetic phenotypes for each dimension may maximally inform initial treatment selection and approach and continue to advance the field of personalized medicine.

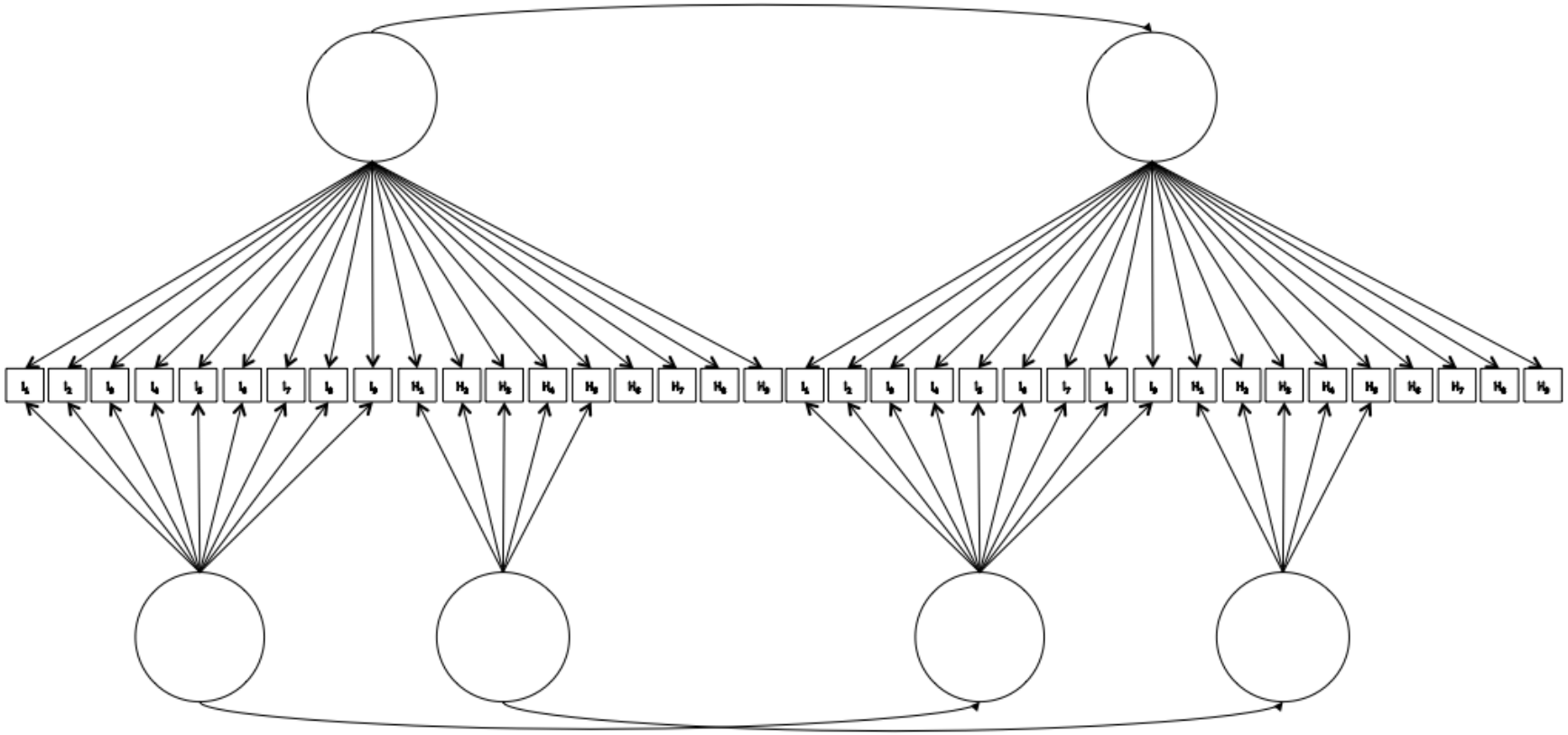

Figure 3-1. Diagram of the multigroup longitudinal restricted bifactor model.

Table 3-1.

Descriptive Statistics for Treatment Arms.

	DMPH (n = 69)		COMBO (n = 70)		Atomoxetine (n = 129)		Placebo (n = 124)	
Male, *n* (%)	46	(66.67)	51	(72.86)	98		103	
Female, *n* (%)	23	(33.33)	19	(27.14)	31		21	
Age, *M* (*SD*)	10.25	(2.00)	10.06	(2.16)	9.7	(1.60)	10	(1.50)
Diagnostic subtypes *n* (%)								
Inattention	34	(49.28)	30	(42.86)	24	(18.60)	24	(19.40)
HI	2	(2.90)	2	(2.86)	1	(.80)	2	(1.60)
Combined	33	(47.83)	38	(54.29)	104	(80.60)	98	(79.00)
Comorbid diagnoses *n* (%)								
ODD	25	(36.23)	17	(24.29)	53	(41.10)	45	(36.30)
Elimination disorders	8	(11.59)	6	(8.57)	10	(7.80)	15	(12.10)
Phobias	18	(26.09)	7	(10.00)	16	(12.40)	13	(10.50)
Dysthymia	2	(2.90)	0		7	(5.40)	5	(4.00)
GAD	6	(8.70)	1	(1.43)	4	(3.10)	3	(2.40)
MDD	0		0		4	(3.10)	4	(3.20)
CGI-S, *M* (*SD*)	4.49	(.53)	4.41	(.50)	4.9	(.80)	4.9	(.80)

Table 3-2

Covariances and Correlations for Latent Dimensions Across Baseline and Treatment Endpoint

	General T1	General T2	Inattention T1	Inattention T2	Hyperactivity T1	Hyperactivity T2
Atomoxetine						
General T1	.81 (.12)					
General T2	.62	1.42 (.21)				
Inattention T1	0	0	.79 (.12)			
Inattention T2	0	0	.48	1.57 (.24)		
Hyperactivity T1	0	0	0	0	1.14 (.19)	
Hyperactivity T2	0	0	0	0	.81	.72 (.12)
Placebo						
General T1	.64 (.10)					
General T2	.78	2.46 (.39)				
Inattention T1	0	0	.72 (.12)			
Inattention T2	0	0	.45	1.61 (.26)		
Hyperactivity T1	0	0	0	0	1.22 (.21)	
Hyperactivity T2	0	0	0	0	.71	1.02 (.17)
DMPH						
General T1	.70 (.14)					
General T2	.73	.31 (.06)				
Inattention T1	0	0	.89 (.18)			
Inattention T2	0	0	.31	.75 (.15)		
Hyperactivity T1	0	0	0	0	2.42 (.50)	
Hyperactivity T2	0	0	0	0	0.68	1.67 (.36)
COMBO						
General T1	1.12 (.23)					
General T2	.58	1.02 (.21)				
Inattention T1	0	0	1.43 (.29)			
Inattention T2	0	0	.61	.82 (.17)		
Hyperactivity T1	0	0	0	0	2.16 (.46)	
Hyperactivity T2	0	0	0	0	.51	1.55 (.34)

Note. T1=baseline. T2=endpoint.

Table 3-3.

Average Symptoms and Latent Variable Means with Associated Significance Tests for Atomoxetine Treatment Trial.

	Atomoxetine					Placebo								
	n = 102		*n* = 102			*n* = 102		*n* = 102						
	Baseline		Endpoint		Δ	Baseline		Endpoint		Δ	*F*	*df*	*p*	d
	M	SD	M	SD		M	SD	M	SD					
ANOVA														
ADHD-RS-IV														
TOTAL	38.77	(8.52)	21.54	(13.09)	-17.23	39.27	(8.11)	32.98	(14.68)	-6.29	36.93	1,192	<.001	.87
INATT	21.14	(4.04)	12.51	(7.25)	-8.63	21.86	(3.76)	18.54	(7.17)	-3.32	27.83	1,192	<.001	.77
HYPER/IMP	17.64	(6.38)	9.03	(7.34)	-8.61	17.37	(5.89)	14.43	(8.67)	-2.94	35.97	1,193	<.001	.87
HYPER	9.87	(3.83)	4.67	(4.28)	-5.20	9.69	(3.73)	7.81	(4.99)	-1.88	32.50	1,193	<.001	.82
	M	SE	M	SE		M	SE	M	SE		z			
IRT														
General	-.03	(.09)	-1.26	(.12)	-1.23	-.05	(.09)	-.37	(.17)	-.32	6.12		<.001	.66
Inattention	-.18	(.09)	-1.24	(.13)	-1.06	.03	(.09)	-.39	(.14)	-.42	3.64		<.001	.51
Hyperactivity	.35	(.11)	-.02	(.09)	-.37	.25	(.12)	.02	(.11)	-.23	1.15		.25	.15

Table 3-4.

Average Symptoms and Latent Variable Means with Associated Significance Tests for DMPH or COMBO Treatment Trial.

	DMPH					COMBO								
	n = 60		*n* = 60			*n* = 56		*n* = 56						
	Baseline		Endpoint		Δ	Baseline		Endpoint		Δ	*F*	*df*	*p*	d
	M	SD	M	SD		M	SD	M	SD					
ANOVA														
ADHD-RS-IV														
TOTAL	35.33	(7.98)	20.38	(8.05)	-14.95	36.32	(9.48)	17.95	(9.77)	-9.77	2.57	1, 114	.11	.30
INATT	21.43	(4.21)	12.82	(5.08)	-8.61	20.79	(5.04)	10.70	(5.60)	-5.60	1.38	1, 114	.24	.22
HYPER/IMP	13.90	(6.41)	7.57	(4.59)	-6.33	15.54	(7.08)	7.25	(5.65)	-5.65	2.81	1, 114	.097	.31
HYPER	7.72	(4.14)	3.73	(2.96)	-3.99	8.46	(4.34)	3.44	(3.29)	-3.29	1.58	1, 114	.21	.24
	M	SE	M	SE		M	SE	M	SE		z			
IRT														
General	-.58	(.11)	-1.35	(.08)	-.77	-.48	(.15)	-1.57	(.14)	-1.09	-2.02		.04	.39
Inattention	.16	(.13)	-1.11	(.12)	-1.27	-.06	(.17)	-1.55	(.13)	-1.49	-1.08		.28	.25
Hyperactivity	.20	(.21)	-.29	(.18)	-.49	.06	(.21)	-.61	(.19)	-.67	-.67		.50	.14

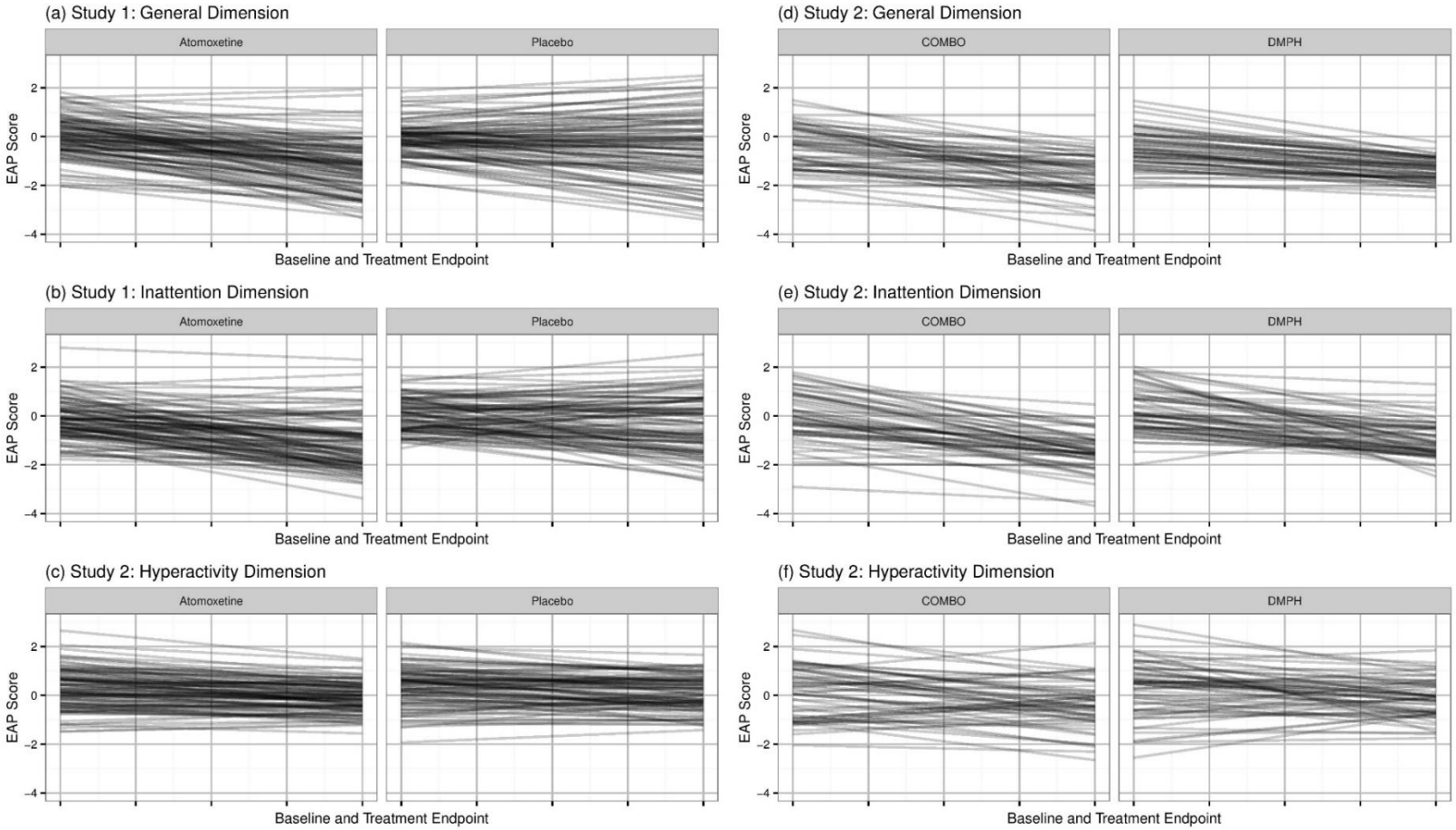

Figure 3-2. Study participant EAP scores for the general, inattention, and hyperactivity subdimensions of each treatment group for baseline and treatment endpoint. Each chart shows the variability in treatment effect across participants.

REFRENCES

Abikoff, H. B., Jensen, P. S., Arnold, L. L. E., Hoza, B., Hechtman, L., Pollack, S., … Wigal, T. (2002). Observed classroom behavior of children with ADHD: Relationship to gender and comorbidity. *Journal of Abnormal Child Psychology*, *30*(4), 349–359. http://doi.org/10.1023/A:1015713807297

Angold, A., Erkanli, A., Egger, H. L., & Costello, E. J. (2000). Stimulant treatment for children: a community perspective. *Journal of the American Academy of Child and Adolescent Psychiatry*, *39*(8), 975-984-994. http://doi.org/10.1097/00004583-200008000-00009

Arnett, A. B., Pennington, B. F., Willcutt, E. G., DeFries, J. C., & Olson, R. K. (2015). Sex differences in ADHD symptom severity. *Journal of Child Psychology and Psychiatry*, *56*(6), 632–639. http://doi.org/10.1111/jcpp.12337

Arnold, L. E. (2000). Methyiphenidate vs. amphetamine: Comparative review. *Journal of Attention Disorders*, *3*(4), 200–211. http://doi.org/10.1177/108705470000300403

Arnsten, A. F., Scahill, L., & Findling, R. L. (2007). alpha2-Adrenergic receptor agonists for the treatment of attention-deficit/hyperactivity disorder: emerging concepts from new data. *Journal of Child and Adolescent Psychopharmacology*, *17*(4), 393–406. http://doi.org/10.1089/cap.2006.0098

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Biederman, J., Krishnan, S., Zhang, Y., McGough, J. J., & Findling, R. L. (2007). Efficacy and tolerability of lisdexamfetamine dimesylate (NRP-104) in children with attention-deficit/hyperactivity disorder: A Phase III, multicenter, randomized, double-blind, forced-dose, parallel-group study. *Clinical Therapeutics*, *29*(3), 450–463. http://doi.org/10.1016/S0149-2918(07)80083-X

Biederman, J., Melmed, R. D., Patel, A., McBurnett, K., Konow, J., Lyne, A., & Scherer, N. (2008). A randomized, double-blind, placebo-controlled study of guanfacine extended release in children and adolescents with attention-deficit/hyperactivity disorder. *Pediatrics*, *121*(1), e73–e84. http://doi.org/10.1542/peds.2006-3695

Biederman, J., Mick, E., & Faraone, S. V. (2000). Age-dependent decline of symptoms of attention deficit hyperactivity disorder: Impact of remission definition and symptom type. *American Journal of Psychiatry*, *157*(5), 816–818. http://doi.org/10.1176/appi.ajp.157.5.816

Biederman, J., Mick, E., Faraone, S. V., Braaten, E., Doyle, A., Spencer, T., … Johnson, M. A. (2002). Influence of gender on attention deficit hyperactivity disorder in children referred to a psychiatric clinic. *American Journal of Psychiatry*, *159*(1), 36–42. http://doi.org/10.1176/appi.ajp.159.1.36

Biederman, J., Monuteaux, M. C., Mick, E., Spencer, T., Wilens, T. E., Silva, J. M., … Faraone, S. V. (2006). Young adult outcome of attention deficit hyperactivity disorder: A controlled 10-year follow-up study. *Psychological Medicine*, *null*(2), 167–179. http://doi.org/10.1017/S0033291705006410

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. http://doi.org/10.1007/BF02293801

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika*, *75*(1), 33–57. http://doi.org/10.1007/s11336-009-9136-x

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*(3), 307–335. http://doi.org/10.3102/1076998609353115

Cai, L. (2012). *flexMIRT: Flexible multilevel item factor analysis and test scoring*. Seattle, WA: Vector Psychometric Group, LLC.

Cai, L., Choi, K., Hansen, M., & Harrell, L. (2016). Item response theory. *Annual Review of Statistics and Its Application*, *3*(1), null. http://doi.org/10.1146/annurev-statistics-041715-033702

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, *16*(3), 221–248. http://doi.org/10.1037/a0023350

Castellanos, F. X., Sonuga-Barke, E. J. S., Milham, M. P., & Tannock, R. (2006). Characterizing cognition in ADHD: beyond executive dysfunction. *Trends in Cognitive Sciences*, *10*(3), 117–123. http://doi.org/10.1016/j.tics.2006.01.011

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., … Rose, M. (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*, *45*(5 Suppl 1), S3–S11. http://doi.org/10.1097/01.mlr.0000258615.42478.55

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289. http://doi.org/10.3102/10769986022003265

Cheng, J. Y. W., Chen, R. Y. L., Ko, J. S. N., & Ng, E. M. L. (2007). Efficacy and safety of atomoxetine for attention-deficit/hyperactivity disorder in children and adolescents-meta-analysis and meta-regression analysis. *Psychopharmacology*, *194*(2), 197–209. http://doi.org/10.1007/s00213-007-0840-x

Childress, A. C. (2012). Guanfacine extended release as adjunctive therapy to psychostimulants in children and adolescents with attention-deficit/hyperactivity disorder. *Advances in Therapy*, *29*(5), 385–400. http://doi.org/10.1007/s12325-012-0020-1

Childress, A. C., & Sallee, F. R. (2014). Attention-deficit/hyperactivity disorder with inadequate response to stimulants: Approaches to management. *CNS Drugs*, *28*(2), 121–129. http://doi.org/10.1007/s40263-013-0130-6

Conners, C. K. (1998). Rating scales in attention-deficit/hyperactivity disorder: use in assessment and treatment monitoring. *The Journal of Clinical Psychiatry*, *59 Suppl 7*, 24–30.

Conners, C. K., Sitarenios, G., Parker, J. D. A., & Epstein, J. N. (1998). The revised Conners' Parent Rating Scale (CPRS-R): Factor structure, reliability, and criterion validity. *Journal of Abnormal Child Psychology*, *26*(4), 257–268. http://doi.org/10.1023/A:1022602400621

Davidson, M. C., Amso, D., Anderson, L. C., & Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia*, *44*(11), 2037–2078. http://doi.org/10.1016/j.neuropsychologia.2006.02.006

DuPaul, G. J., Power, T. J., Anastopoulos, A. D., & Reid, R. (1998). *ADHD Rating Scale-IV: Checklists, Norms, and Clinical Interpretations*. New York, NY: The Guilford Press.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Faraone, S. V., Biederman, J., Spencer, T. J., & Aleardi, M. (2006). Comparing the efficacy of medications for ADHD using meta-analysis. *Medscape General Medicine*, *8*(4), 4.

Faraone, S. V., & Buitelaar, J. (2009). Comparing the efficacy of stimulants for ADHD in children and adolescents using meta-analysis. *European Child & Adolescent Psychiatry*, *19*(4), 353–364. http://doi.org/10.1007/s00787-009-0054-3

Froehlich, T. E., Lanphear, B. P., Epstein, J. N., Barbaresi, W. J., Katusic, S. K., & Kahn, R. S. (2007). Prevalence, recognition, and treatment of attention-deficit/hyperactivity disorder in a national sample of us children. *Archives of Pediatrics & Adolescent Medicine*, *161*(9), 857–864. http://doi.org/10.1001/archpedi.161.9.857

Garcia-Rosales, A., Vitoratou, S., Banaschewski, T., Asherson, P., Buitelaar, J., Oades, R. D., … Chen, W. (2015). Are all the 18 DSM-IV and DSM-5 criteria equally useful for diagnosing ADHD and predicting comorbid conduct problems? *European Child & Adolescent Psychiatry*, *24*(11), 1325–1337. http://doi.org/10.1007/s00787-015-0683-7

Gershon, J., & Gershon, J. (2002). A meta-analytic review of gender differences in ADHD. *Journal of Attention Disorders*, *5*(3), 143–154. http://doi.org/10.1177/108705470200500302

Gibbins, C., Toplak, M. E., Flora, D. B., Weiss, M. D., & Tannock, R. (2012). Evidence for a general factor model of ADHD in adults. *Journal of Attention Disorders*, *16*(8), 635–644. http://doi.org/10.1177/1087054711416310

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*(3), 423–436. http://doi.org/10.1007/BF02295430

Gomez, R. (2008). Item response theory analyses of the parent and teacher ratings of the DSM-IV ADHD Rating Scale. *Journal of Abnormal Child Psychology*, *36*(6), 865–885. http://doi.org/10.1007/s10802-008-9218-8

Gomez, R., Vance, A., & Gomez, R. M. (2013). Validity of the ADHD bifactor model in general community samples of adolescents and adults, and a clinic-referred sample of children and adolescents. *Journal of Attention Disorders*. http://doi.org/10.1177/1087054713480034

Hansen, M., Cai, L., Stucky, B. D., Tucker, J. S., Shadel, W. G., & Edelen, M. O. (2014). Methodology for developing and evaluating the PROMIS smoking item banks. *Nicotine & Tobacco Research: Official Journal of the Society for Research on Nicotine and Tobacco*, *16 Suppl 3*, S175-189. http://doi.org/10.1093/ntr/ntt123

Harpin, V., Mazzone, L., Raynaud, J. P., Kahle, J., & Hodgkins, P. (2016). Long-term outcomes of ADHD: A systematic review of self-esteem and social function. *Journal of Attention Disorders*, *20*(4), 295–305. http://doi.org/10.1177/1087054713486516

Hart, E. L., Lahey, B. B., Loeber, R., Applegate, B., & Frick, P. J. (1995). Developmental change in attention-deficit hyperactivity disorder in boys: A four-year longitudinal study. *Journal of Abnormal Child Psychology*, *23*(6), 729–749.

Hinshaw, S. P. (2007). Moderators and mediators of treatment outcome for youth with ADHD: Understanding for whom and how interventions work. *Ambulatory Pediatrics*, *7*(1, Supplement), 91–100. http://doi.org/10.1016/j.ambp.2006.04.012

Hirota, T., Schwartz, S., & Correll, C. U. (2014). alpha-2 agonists for attention-deficit/hyperactivity disorder in youth: A systematic review and meta-analysis of monotherapy and add-on trials to stimulant therapy. *Journal of the American Academy of Child & Adolescent Psychiatry*, *53*(2), 153–173. http://doi.org/10.1016/j.jaac.2013.11.009

Insel, T. (2015, February 2). Director's blog: Precision medicine for mental disorders. Retrieved from http://www.nimh.nih.gov/about/director/2015/precision-medicine-for-mental-disorders.shtml

Jain, R., Segal, S., Kollins, S. H., & Khayrallah, M. (2011). Clonidine extended-release tablets for pediatric patients with attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, *50*(2), 171–179. http://doi.org/10.1016/j.jaac.2010.11.005

Jensen, P. S., Kettle, L., Roper, M. T., Sloan, M. T., Dulcan, M. K., Hoven, C., … Payne, J. D. (1999). Are stimulants overprescribed? Treatment of ADHD in four U.S. communities. *Journal of the American Academy of Child and Adolescent Psychiatry*, *38*(7), 797–804. http://doi.org/10.1097/00004583-199907000-00008

Kaufman, J., Birmaher, B., Brent, D., Rao, U., Flynn, C., Moreci, P., … 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 & Adolescent Psychiatry*, *36*(7), 980–988. http://doi.org/10.1097/00004583-199707000-00021

Lahey, B. B., Hartung, C. M., Loney, J., Pelham, W. E., Chronis, A. M., & Lee, S. S. (2007). Are there sex differences in the predictive validity of DSM-IV ADHD among younger children? *Journal of Clinical Child and Adolescent Psychology*, *36*(2), 113–126. http://doi.org/10.1080/15374410701274066

Lahey, B. B., Pelham, W. E., Loney, J., Lee, S. S., & Willcutt, E. (2005). Instability of the DSM-IV subtypes of ADHD from preschool through elementary school. *Archives of General Psychiatry*, *62*(8), 896–902. http://doi.org/10.1001/archpsyc.62.8.896

Langer, M. M. (2008). *A reexamination of Lord's Wald test for differential item functioning using item response theory and modern error estimation*. The University of North Carolina at Chapel Hill. Retrieved from http://gradworks.umi.com/33/31/3331000.html

Martel, M. M., Eye, A. von, & Nigg, J. (2012). Developmental differences in structure of attention-deficit/hyperactivity disorder (ADHD) between childhood and adulthood. *International Journal of Behavioral Development*, *36*(4), 279–292. http://doi.org/10.1177/0165025412444077

Martel, M. M., Roberts, B., Gremillion, M., von Eye, A., & Nigg, J. T. (2011). External validation of bifactor model of ADHD: Explaining heterogeneity in psychiatric comorbidity, cognitive control, and personality trait profiles within DSM-IV ADHD. *Journal of Abnormal Child Psychology*, *39*(8), 1111–1123. http://doi.org/10.1007/s10802-011-9538-y

Martel, M. M., von Eye, A., & Nigg, J. T. (2010). Revisiting the latent structure of ADHD: is there a "g" factor? *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *51*(8), 905–914. http://doi.org/10.1111/j.1469-7610.2010.02232.x

Maydeu-Olivares, A. (2013). Why should we assess the goodness-of-fit of IRT models? *Measurement: Interdisciplinary Research and Perspectives*, *11*(3), 127–137. http://doi.org/10.1080/15366367.2013.841511

McArdle, J. . J., Petway, K. . T., & Hishinuma, E. S. (2015). IRT for growth and change. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 435–456). New York: Routledge.

McCracken, J. T., McGough, J. J., Loo, S. K., Levitt, J., Del'Homme, M., Cowen, J., … Bilder, R. (under review). Combined stimulant and guanfacine administration in attention-deficit hyperactivity disorder (ADHD): A controlled, comparative study.

McGough, J. J., McCracken, J. T., Swanson, J. M., Riddle, M. A., Kollins, S. H., Greenhill, L., … Vitiello, B. (2006). Pharmacogenetics of methylphenidate response in preschoolers with ADHD. *Journal of the American Academy of Child & Adolescent Psychiatry*, *45*(11), 1314–1322. http://doi.org/10.1097/01.chi.0000235083.40285.08

Merikangas, K. R., He, J.-P., Brody, D., Fisher, P. W., Bourdon, K., & Koretz, D. S. (2010). Prevalence and treatment of mental disorders among U.S. children in the 2001–2004 NHANES. *Pediatrics*, *125*(1), 75–81. http://doi.org/10.1542/peds.2008-2598

Michelson, D., Allen, A. J., Busner, J., Casat, C., Dunn, D., Kratochvil, C., … Harder, D. (2002). Once-daily atomoxetine treatment for children and adolescents with attention deficit hyperactivity

disorder: a randomized, placebo-controlled study. *The American Journal of Psychiatry*, *159*(11), 1896–1901. http://doi.org/10.1176/appi.ajp.159.11.1896

Michelson, D., Faries, D., Wernicke, J., Kelsey, D., Kendrick, K., Sallee, F. R., … Group, the A. A. S. (2001). Atomoxetine in the treatment of children and adolescents with attention-deficit/hyperactivity disorder: A randomized, placebo-controlled, dose-response study. *Pediatrics*, *108*(5), e83–e83. http://doi.org/10.1542/peds.108.5.e83

Mislevy, D. R. J. (1984). Estimating latent distributions. *Psychometrika*, *49*(3), 359–381. http://doi.org/10.1007/BF02306026

Molina, B. S. G., Hinshaw, S. P., Eugene Arnold, L., Swanson, J. M., Pelham, W. E., Hechtman, L., … MTA Cooperative Group. (2013). Adolescent substance use in the multimodal treatment study of attention-deficit/hyperactivity disorder (ADHD) (MTA) as a function of childhood ADHD, random assignment to childhood treatments, and subsequent medication. *Journal of the American Academy of Child and Adolescent Psychiatry*, *52*(3), 250–263. http://doi.org/10.1016/j.jaac.2012.12.014

Monroe, S., & Cai, L. (2015). Evaluating structural equation models for categorical outcomes: A new test statistic and a practical challenge of interpretation. *Multivariate Behavioral Research*, *50*(6), 569–583. http://doi.org/10.1080/00273171.2015.1032398

National Institutes of Health. (2003). Final NIH statement on sharing research data. Retrieved from http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html

Normand, S., Flora, D. B., Toplak, M. E., & Tannock, R. (2012). Evidence for a general ADHD factor from a longitudinal general school population study. *Journal of Abnormal Child Psychology*, *40*(4), 555–567. http://doi.org/10.1007/s10802-011-9584-5

Pliszka, S. (2007). Practice parameter for the assessment and treatment of children and adolescents with attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, *46*(7), 894–921. http://doi.org/10.1097/chi.0b013e318054e724

Pliszka, S. R. (2012). Psychostimulants. In D. R. Rosenberg & S. Gershon (Eds.), *Pharmacotherapy of Child and Adolescent Psychiatric Disorders* (pp. 65–104). John Wiley & Sons, Ltd. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9781119958338.ch6/summary

Polanczyk, G., de Lima, M. S., Horta, B. L., Biederman, J., & Rohde, L. A. (2007). The worldwide prevalence of ADHD: A systematic review and metaregression analysis. *American Journal of Psychiatry*, *164*(6), 942–948. http://doi.org/10.1176/ajp.2007.164.6.942

Purpura, D. J., Wilson, S. B., & Lonigan, C. J. (2010). Attention-deficit/hyperactivity disorder symptoms in preschool children: Examining psychometric properties using item response theory. *Psychological Assessment*, *22*(3), 546–558. http://doi.org/http://dx.doi.org/10.1037/a0019581

Reckase, M. (2009). *Multidimensional item response theory (Vol. 150)*. New York: Springer.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, *16*(1), 19–31. http://doi.org/10.1007/s11136-007-9183-7

Riddle, M. A., Yershova, K., Lazzaretto, D., Paykina, N., Yenokyan, G., Greenhill, L., … Posner, K. (2013). The preschool attention-deficit/hyperactivity disorder treatment study (PATS) 6-Year follow-up. *Journal of the American Academy of Child & Adolescent Psychiatry*, *52*(3), 264–278.e2. http://doi.org/10.1016/j.jaac.2012.12.007

Rucklidge, J. J. (2010). Gender differences in attention-deficit/hyperactivity disorder. *Psychiatric Clinics of North America*, *33*(2), 357–373. http://doi.org/10.1016/j.psc.2010.01.006

Sallee, F. R., Lyne, A., Wigal, T., & McGough, J. J. (2009). Long-term safety and efficacy of guanfacine extended release in children and adolescents with attention-deficit/hyperactivity disorder. *Journal of Child and Adolescent Psychopharmacology*, *19*(3), 215–226. http://doi.org/10.1089/cap.2008.0080

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*(4, Pt. 2), 100.

Scahill, L., Chappell, P. B., Kim, Y. S., Schultz, R. T., Katsovich, L., Shepherd, E., … Leckman, J. F. (2001). A placebo-controlled study of guanfacine in the treatment of children with tic disorders and attention deficit hyperactivity disorder. *American Journal of Psychiatry*, *158*(7), 1067–1074. http://doi.org/10.1176/appi.ajp.158.7.1067

Shaw, M., Hodgkins, P., Caci, H., Young, S., Kahle, J., Woods, A. G., & Arnold, L. E. (2012). A systematic review and analysis of long-term outcomes in attention deficit hyperactivity disorder: effects of treatment and non-treatment. *BMC Medicine*, *10*, 99. http://doi.org/10.1186/1741-7015-10-99

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*(3), 237–247.

Smith, L. C., Tamm, L., Hughes, C. W., & Bernstein, I. H. (2013). Separate and overlapping relationships of inattention and hyperactivity/impulsivity in children and adolescents with attention-deficit/hyperactivity disorder. *Attention Deficit and Hyperactivity Disorders*, *5*(1), 9–20. http://doi.org/10.1007/s12402-012-0091-5

Solanto, M., Newcorn, J., Vail, L., Gilbert, S., Ivanov, I., & Lara, R. (2009). Stimulant drug response in the predominantly inattentive and combined subtypes of attention-deficit/hyperactivity disorder. *Journal of Child and Adolescent Psychopharmacology*, *19*(6), 663–671. http://doi.org/10.1089/cap.2009.0033

Spencer, T., Heiligenstein, J. H., Biederman, J., Faries, D. E., Kratochvil, C. J., Conners, C. K., & Potter, W. Z. (2002). Results from 2 proof-of-concept, placebo-controlled studies of atomoxetine in children with attention-deficit/hyperactivity disorder. *The Journal of Clinical Psychiatry*, *63*(12), 1140–1147.

Stucky, B. D., & Edelen, M. O. (2015). Using hierarchical IRT models to create unidimensional measures from multidimensional data. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling* (pp. 183–206). New York, NY: Routledge.

Swanson, J. M. (1992). *School-based Assessments and Interventions for ADD Students*. K.C. Publishing.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement*, *26*(3), 247–260. http://doi.org/10.1111/j.1745-3984.1989.tb00331.x

Thissen, D., & Wainer, H. (2001). *Test Scoring*. New Jersey: Lawrence Erlbaum Associates, Inc.

Toplak, M. E., Pitch, A., Flora, D. B., Iwenofu, L., Ghelani, K., Jain, U., & Tannock, R. (2009). The unity and diversity of inattention and hyperactivity/impulsivity in ADHD: evidence for a general factor with separable dimensions. *Journal of Abnormal Child Psychology*, *37*(8), 1137–1150. http://doi.org/10.1007/s10802-009-9336-y

Toplak, M. E., Sorge, G. B., Flora, D. B., Chen, W., Banaschewski, T., Buitelaar, J., … Faraone, S. V. (2012). The hierarchical factor model of ADHD: invariant across age and national groupings? *Journal of Child Psychology and Psychiatry*, *53*(3), 292–303. http://doi.org/10.1111/j.1469-7610.2011.02500.x

Toplak, M. E., West, R. F., & Stanovich, K. E. (2012). Do performance-based measures and ratings of executive function assess the same construct? *Journal of Child Psychology and Psychiatry*, no–no. http://doi.org/10.1111/jcpp.12001

Ullebø, A. K., Breivik, K., Gillberg, C., Lundervold, A. J., & Posserud, M.-B. (2012). The factor structure of ADHD in a general population of primary school children. *Journal of Child Psychology and Psychiatry*, *53*(9), 927–936. http://doi.org/10.1111/j.1469-7610.2012.02549.x

Wagner, F., Martel, M. M., Cogo-Moreira, H., Maia, C. R. M., Pan, P. M., Rohde, L. A., & Salum, G. A. (2016). Attention-deficit/hyperactivity disorder dimensionality: the reliable "g" and the elusive "s" dimensions. *European Child & Adolescent Psychiatry*, *25*(1), 83–90. http://doi.org/10.1007/s00787-015-0709-1

Wigal, S. B. (2012). Efficacy and safety limitations of attention-deficit hyperactivity disorder pharmacotherapy in children and adults. *CNS Drugs*, *23*(1), 21–31. http://doi.org/10.2165/00023210-200923000-00004

Willcutt, E. G. (2012). The prevalence of DSM-IV attention-deficit/hyperactivity disorder: A meta-analytic review. *Neurotherapeutics*, *9*(3), 490–499. http://doi.org/10.1007/s13311-012-0135-8

Willcutt, E. G., Doyle, A. E., Nigg, J. T., Faraone, S. V., & Pennington, B. F. (2005). Validity of the executive function theory of attention-deficit/hyperactivity disorder: A meta-analytic review. *Biological Psychiatry*, *57*(11), 1336–1346. http://doi.org/10.1016/j.biopsych.2005.02.006

Willcutt, E. G., Nigg, J. T., Pennington, B. F., Solanto, M. V., Rohde, L. A., Tannock, R., … Lahey, B. B. (2012). Validity of DSM-IV attention deficit/hyperactivity disorder symptom dimensions and subtypes. *Journal of Abnormal Psychology*, *121*(4), 991–1010. http://doi.org/http://dx.doi.org/10.1037/a0027347

Willoughby, M. T., Blanton, Z. E., & Investigators, F. L. P. (2015). Replication and external validation of a bi-factor parameterization of attention deficit/hyperactivity symptomatology. *Journal of Clinical Child & Adolescent Psychology*, *44*(1), 68–79. http://doi.org/10.1080/15374416.2013.850702

Wolraich, M. L., Greenhill, L. L., Pelham, W., Swanson, J., Wilens, T., Palumbo, D., … Group, on behalf of the C. S. (2001). Randomized, Controlled Trial of OROS Methylphenidate Once a Day in Children With Attention-Deficit/Hyperactivity Disorder. *Pediatrics*, *108*(4), 883–892. http://doi.org/10.1542/peds.108.4.883

Wolraich, M. L., Hannah, J. N., Baumgaertel, A., & Feurer, I. D. (1998). Examination of DSM-IV criteria for attention deficit/hyperactivity disorder in a county-wide sample. *Journal of Developmental and Behavioral Pediatrics: JDBP*, *19*(3), 162–168.

Woods, C. M. (2007). Empirical histograms in item response theory with ordinal data. *Educational and Psychological Measurement*, *67*(1), 73–87. http://doi.org/10.1177/0013164406288163

Woods, C. M., Cai, L., & Wang, M. (2013). The Langer-Improved Wald test for DIF testing with multiple groups evaluation and comparison to two-group IRT. *Educational and Psychological Measurement*, *73*(3), 532–547. http://doi.org/10.1177/0013164412464875