DISSERTATION

# RASCH ANALYSIS OF THE EVALUATION IN AYRES SENSORY INTEGRATION (EASI)

Submitted by

Patricia Grady

Department of Occupational Therapy

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2022

Doctoral Committee

Advisor: Anita Bundy

Shelly Lane
Susan Hepburn
Julia Sharp

# ABSTRACT

## RASCH ANALYSIS OF THE EVALUATION IN AYRES SENSORY INTEGRATION (EASI)

Sensory Integration (SI) refers to the neurological process by which a person takes in sensory information, interprets this information, and uses it to inform movement and goal-directed action (Ayres, 1989). For children with a variety of diagnoses, as well as some children who are otherwise typically developing, SI may be impaired or delayed (Bundy & Lane, 2020a). Occupational therapists and other clinicians who treat children with sensory integration (SI) dysfunction face a dearth of appropriate instruments to evaluate SI function.

The Evaluation in Ayres Sensory Integration (EASI; Mailloux et al., 2018) is a novel assessment tool that may allow therapists to evaluate SI in a way that is aligned with SI theory. EASI consists of 21 individual tests that measure constructs of SI. The EASI authors have collected normative data for children across the globe. However, this data must be assessed for validity and reliability before it can be used as the basis for normative scoring on the EASI.

In this dissertation, I used Rasch analysis to evaluate data with 19 of the 21 tests. The Rasch model is a latent-trait psychometric model that (1) transforms ordinal-level data to interval-level data and (2) allows users to assess evidence for construct validity (unidimensionality and invariance) and internal reliability (Bond, Yan & Heene, 2020). For each of the 19 tests, I evaluated item fit statistics, rating scale fit statistics, person fit statistics, principal components analysis (PCA) of standardized residuals, differential item functioning (DIF) based on sex, person reliability index and strata. The dataset for this study comprised 2653 children from 51 countries; all data were collected by trained EASI examiners.

Overall, results revealed promising evidence for construct validity and internal reliability of data collected using 16 of the19 EASI tests evaluated in this dissertation. However, across many tests, I observed lower-than-desired person fit statistics and reliability. Notably, these results were not far from the desired values. I hypothesized that these findings are the result of the overall high ability level of the normative population. EASI is designed to evaluate children with lower-than-average or poor SI function; therefore, these findings are not unexpected nor are they particularly concerning.

Three EASI tests (Proprioception: Force, Proprioception: Joint Position, and Balance) required substantial revision as a result of these analyses; each had threats to construct validity that exceeded my expectations. In this dissertation, I provided potential solutions for these three tests; future studies will evaluate the extent to which these solutions resolve concerns.

In conclusion, the normative data form an acceptable basis for creating norm-based scores for clinical interpretation. However, larger studies must be conducted with clinical populations to ensure that the tests can differentiate between children with and without SI dysfunction. Further, future studies should investigate the role of culture, language and other factors on the validity of EASI test scores.

ACKNOWLEDGEMENTS

It is my pleasure to thank the members of my dissertation committee for their support during my dissertation process. Dr. Susan Hepburn offered excellent insight into the strengths and challenges of children with developmental disabilities. Dr. Julia Sharp strengthened this dissertation through her thoughtful statistical feedback. Dr. Shelly Lane provided wisdom in SI theory as well as neurological underpinnings and intervention/assessment approaches.

I am eternally grateful to my committee chair, Dr. Anita Bundy. This dissertation would not have been possible without her mentorship. She patiently taught me Rasch analysis, guided me through SI theory, and helped me navigate the research and writing process. In addition, she has always recognized me as a whole person – she has been understanding and supportive when other aspects of my life intruded on our work. She has been patient and kind, offering humor and encouragement. I am the researcher that I am because of her, but I am also the person that I am because of her.

I am thankful, too, for the support from the College of Health and Human Science's Dean's Office. The Dean's Fellowship grant that I received in 2021 financially supported this project. Thank you for believing in the value of my work; this support gave me not only the funding and dedicated time to work on my dissertation, but also the confidence.

I also thank Dr. Zoe Mailloux, Dr. Diane Parham, Dr. Susanne Smith Roley, Dr. Roseann Schaaf, and the entire Collaboration for Leadership in Ayres Sensory Integration (CLASI) team. In addition to sharing data, you have all scaffolded my work and helped me grow as a researcher. I am especially thankful to Dr. Mailloux; her determination to create the best EASI possible has underscored my own dedication.

I would also like to acknowledge the entire Colorado State University – Occupational Therapy faculty, staff, and students who, throughout the last six years, have watched me grow into an occupational therapist and researcher. Debi Krogh-Michna, the candy bowl in your office fueled this dissertation (your warmth and kindness did not hurt it either). I am grateful to my graduate research assistant, Katya Navrotskaya, for her efforts in this project. Thank you to all the PhD students who have been along with me on this ride. I cannot wait to see what you do next.

Last but far from least, I would like to thank my family and friends for their support during my PhD process. To my husband, Dennis – you have unwaveringly and unselfishly stood by me through the most difficult and most wonderful moments of my life. My accomplishments are yours as well. To my mother – you have challenged me, cheered for me, and encouraged me since childhood. To my father – you have always believed in my ability to accomplish great things. To my sister, Caroline – you have laughed with me, cried with me, and shown me how to break down any barrier to meet my goals. Kori – you have been more than a best friend, but also a part of my family. You have taught me confidence and grace that have been invaluable as I have grown in my career. To my amazing friends, and especially the community of parent friends who have supported me through pandemic parenting, I would not be here, finishing my PhD, without your faith, assistance, and advice. My gratitude is not limited to those listed here; countless individuals have made this journey possible. I thank you all.

# DEDICATION

This dissertation is dedicated to my children. Luna, you are the sunshine of my brightest days and the moonlight on my darkest nights. You are a joy to raise. Julian, although I have not seen your little face yet, you are my motivation, my hope for the future, and my constant companion. I dreamed of you both long before I knew you. I hope I have made you proud.

TABLE OF CONTENTS

# CHAPTER 1: INTRODUCTION

Sensory integration (SI) refers to "the neurological process that organizes sensations from one's body and from the environment and makes it possible to use the body effectively in the environment" (Ayres, 1989, p. 11). Occupational therapist A. Jean Ayres introduced her theory of sensory integration in the mid-20th century. In the decades since the conception of SI theory, many scholars in the disciplines of occupational therapy and occupational science have built upon her work. SI theory is both a basic and an applied theory (Bundy & Lane, 2020b). The basic theory examines function and dysfunction in SI, while the applied theory gives rise to assessments and interventions to treat SI dysfunction.

## Function and Dysfunction in Sensory Integration

According to Ayres' (1972) early conceptualization of sensory integration, integration of input from the vestibular, proprioceptive, and tactile systems allows basic motor outputs, such as eye movement, posture, and balance. These outputs allow for more advanced integration, which in turn facilitates more complex functions, such as praxis (complex or novel motor planning) and bilateral coordination. Visual and auditory inputs, likewise, contribute to increasingly more complex behaviors such as maintenance of attention, language development, and hand-eye coordination. These functional skills enable higher-order behaviors such as concentration, self-esteem, and capacity for reasoning among other foundations for new learning. SI gives rise to learning in a cyclical pattern, where sensory intake leads to integration, which in turn leads to planning and organizing behavior and enacting the behavior through an *adaptive response* to the sensory intake (Bundy & Lane, 2020b). Bundy and Lane (2020b) defined an adaptive response as a successful, purposeful, goal-directed action. The adaptive response leads to new sensory

feedback, and the cycle continues. According to Ayres (1972), the child seeks out opportunities to attempt more and more sophisticated adaptive responses; she termed this acting on *inner drive.*

In children with a variety of clinical diagnoses, as well as some children who are otherwise typically developing, sensory integration may be impaired or delayed. Estimates suggest that as many as 1 in every 10 children experiences SI difficulties (Ben-Sasson, Carter, & Briggs-Gowan, 2009; Ahn et al., 2004). Based on a long tradition of research conducted by Ayres and her successors, Bundy and Lane (2020a) created a model of SI dysfunction. They described two overarching categories of dysfunction: poor sensory modulation and poor sensory integration and praxis (dyspraxia). Figure 1 displays the Bundy and Lane (2020a) model, which connects discrete sensory modalities with patterns of dysfunction and downstream behavioral consequences, as well as the brain regions thought to be involved. This model is largely based upon Ayres' work; it expands the role of sensory modulation and more discretely defines brain regions involved in these processes.

Poor sensory modulation (left side of Figure 1) may manifest as over-responsivity or under-responsivity to sensory input (Bundy & Lane, 2020a). Over-responsivity presents either as an unexpected degree of sensory sensitivity or as avoidance; the child may react to the sensation in a way that seems out-of-proportion, or they may avoid the sensation entirely. Under-responsivity is characterized as a greatly diminished response; the child may seem not to notice the stimuli. Some children present with fluctuating patterns of responsivity in different contexts or to different types of stimuli.

The other pattern of dysfunction in the Bundy and Lane (2020a) model describes children with overall poor sensory integration, resulting in difficulties with motor function and motor

planning (i.e., praxis; right side of Figure 1). Dyspraxia refers to difficulty conceptualizing, planning and executing movements. It also includes poor *ideation*, the process of identifying and understanding the affordances of objects and environment (i.e., knowing what actions can be done in a given situation). Bundy and Lane (2020a) further divided the dyspraxia construct into two conditions: vestibular bilateral integration and sequencing dysfunction (VBIS) and somatodyspraxia. VBIS presents as difficulty with bilateral coordination and anticipatory "feed-forward" movements (those that require anticipation of a future event; e.g., kicking a ball that is in motion). VBIS generally stems from poor postural and ocular control because of deficits in processing vestibular and proprioceptive input. Somatodyspraxia, the more severe form of dyspraxia, stems from poor body scheme as a result of deficits in vestibular, proprioceptive, and tactile discrimination in conjunction with general difficulty discriminating characteristics of external sensory input (e.g., visual). Children with somatodyspraxia struggle to execute both feed-forward movements and simpler movements that only require the child to integrate sensory input from the environment before acting (e.g., kicking a ball that is not in motion).

As suggested on the far left and far right sides of Figure 1, children with SI deficits may present with a variety of occupational challenges (Parham & Cosbey, 2020). Children with over-responsivity to tactile sensory input, for example, may struggle to wash their skin or hair during bathing. Children with VBIS may be left out of ball play with their classmates on the playground. Children with somatodyspraxia may have trouble dressing themselves. As these children enter middle childhood and adolescence, they may lag behind their peers in education, social relationships, and other key developmental areas (Parham & Cosbey, 2020).

## Intervention using Sensory Integration Theory

Ayres' (1979) theory focused not only on function and dysfunction in SI networks, but

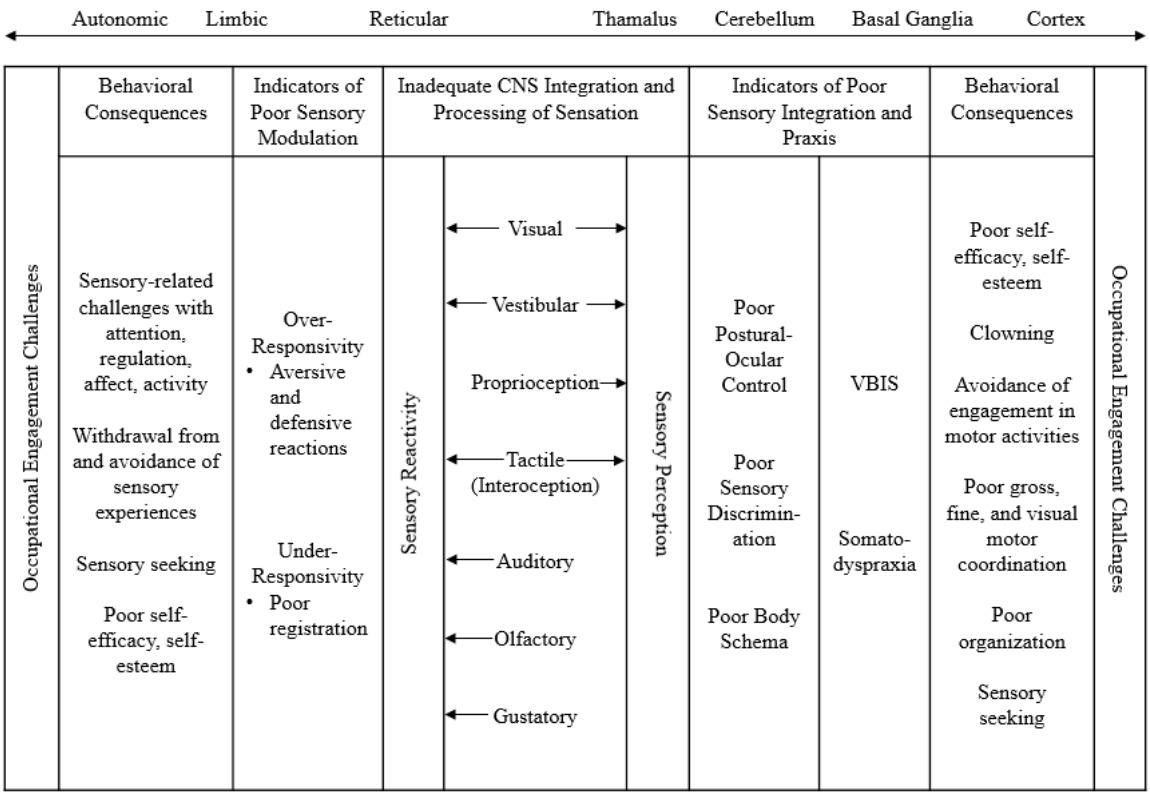



**Figure 1.1**

*SI Dysfunction Schematic Adapted from Bundy & Lane (2020a).* CNS = central nervous system, VBIS = vestibular bilateral integration and sequencing deficit.

also on occupational therapy intervention to improve SI. During SI intervention, the therapist collaborates with the child to identify play-based sensorimotor activities that present a "just-right challenge" to the child's nervous system, prompting an adaptive response (Ayres, 1979; Parham et al., 2011). By taking advantage of the child's inner drive to explore and engage in more complex adaptive responses, the therapist promotes neuroplasticity that may enhance sensory integrative function and repair dysfunction (Lane, 2020).

Over time, several systematic reviews have found mixed evidence for the efficacy of SI intervention. Some of these studies demonstrated promising results (May-Benson & Koomar, 2010; Ottenbacher, 1982; Schaaf et al., 2018; Vargas & Camilli, 1999). These reviews suggested

that SI intervention may positively impact children's motor skills, social behavior, individual functional goals, and participation in occupations of childhood. Other reviews, however, have called into question the efficacy of these interventions (Hoehn & Baumeister, 1994; Leong, Carter & Stephenson, 2015; Polatajko, Kaplan & Wilson, 1992), finding no benefit over traditional occupational therapy.

The varied results across systematic reviews may be the result of unstandardized application of SI intervention (Parham et al., 2011; Schaaf et al., 2018). Parham and colleagues (2011) developed a fidelity measure to assess adherence to the principles of Ayres' SI therapy: the Ayres Sensory Integration Fidelity Measure (ASIFM). In Schaaf and colleagues' (2018) systematic review of SI intervention studies, the authors found only five studies that adhered to these principles; all these studies focused on children with ASD. Among these studies, they found strong evidence suggesting that SI intervention can lead to improved occupational function and participation based on individual goals (i.e., goal attainment scaling). Notably, none of these studies were included in the three reviews finding no benefit of SI intervention.

## Assessment in Sensory Integration

Inconsistent results across intervention studies may be related to the lack of feasible and thorough assessment tools for SI. Assessment of SI function is a cornerstone of effective SI intervention; indeed, it is among the core structural components of intervention described in ASIFM (Parham et al., 2011). In Ayres' SI therapy, therapeutic play sessions must be tailored to the child's individual profile of sensory function. Therapists must conduct thorough assessment to identify "meaningful clusters" of behaviors that indicate underlying sensory function/dysfunction (Bundy & Lane, 2020b; Mulligan, 2020). Therefore, therapists must have appropriate, valid, and reliable tools to evaluate these functions as a basis for intervention.

Currently, the Sensory Integration and Praxis Tests (SIPT; Ayres, 2005) serve as the gold standard for SI evaluation. This suite of 17 tests, developed in the 1960s, 70s, and 80s, measures tactile, proprioceptive, vestibular, and visual perception as well as sensorimotor functions including visual motor integration, bilateral integration, and praxis. However, SIPT have several important limitations. First, they were developed and normed over 40 years ago, and the normative sample only included children from North America. Therefore, the generalizability of these norms is questionable. Second, SIPT omit several aspects of sensory integration now recognized as important, including sensory modulation (also called sensory reactivity in some literature), auditory perception, and ideational praxis (i.e., the ability to generate novel ideas). Third, evidence for some aspects of validity and reliability of data collected using SIPT is questionable (see Chapter 2, Literature *Review* for more details). Finally, and perhaps most importantly, SIPT require extensive time and training, and they are costly to administer and score. In addition to the training and materials required to give SIPT, therapists must use a proprietary scoring program to calculate results (approximately $37 USD per score report). As a result, SIPT are inaccessible to many underfunded clinics both within the US and internationally.

In lieu of SIPT, therapists must rely on parent/caregiver-report instruments such as Sensory Profile 2 (Dunn, 2014) or performance-based developmental instruments that are not grounded in SI theory (e.g., BOT-2, Bruininks & Bruininks, 2005). Like SIPT, these instruments have important limitations (see Chapter 2, *Literature Review* for a full review of common instruments used to evaluate SI functions). Many are not grounded in SI theory, and most are not available or appropriately normed for international populations. Furthermore, no currently available instrument covers all the constructs of SI theory. Therefore, a novel instrument that

produces valid and reliable data is a crucial step for the development and evaluation of SI theory and SI intervention.

## Evaluation in Ayres Sensory Integration (EASI)

Recognizing the shortcomings of existing assessment tools, Mailloux and colleagues (2018) developed a novel approach for assessment of SI function. The Evaluation in Ayres Sensory Integration (EASI) is a new suite of tests that provides clinician-rated, performance-based data about children's sensory integration function (Mailloux et al., 2018). The 21 tests that comprise EASI (described in Table 1) evaluate each component of sensory integration. The tests are grounded in, and reflect, the most current understandings of Ayres' SI theory; they present the most comprehensive operationalization of the theory to date (Mailloux et al., 2018). Unlike SIPT, EASI contains measures of sensory reactivity, auditory perception, and ideational praxis. Furthermore, EASI is much more cost effective than SIPT – most materials are easily purchased from local retailers, and the few standardized materials can be 3D printed for less than $100 USD. The scoring program will be free and available online for immediate results.

The EASI measures four broad constructs involved in SI function: sensory perception, praxis, ocular/postural/bilateral motor integration, and sensory reactivity (called modulation in the Bundy and Lane [2020a] model). EASI addresses multiple sensory modalities, including auditory, tactile, vestibular, proprioceptive, and olfactory. Figure 2 maps the 21 EASI tests onto an abridged version of the Bundy and Lane (2020a) model.

Currently, international normative data collection is underway for the EASI tests in 18 different languages and 81 countries. This suite of instruments may represent a new gold standard for the evaluation of SI functions. Supplementary File 1 contains the full suite of EASI tests.

**Table 1.1**

*Description of EASI Tests*

| # | Test | Brief Description | Items | Item Types | Scoring |
|---|---|---|---|---|---|
| 1 | Visual Praxis: Designs (VPrD) | Participant copies two-dimensional designs; some items have a dot grid while others are free-handed | 24[1] | Accuracy | 2 – 0 (correct/approximate/incorrect) |
| 2 | Visual Perception: Search (VPS) | Participant locates a visual stimulus on one of 3 visually crowded forms within 10, 20 or 30 seconds (dependent on form) | 18 | Accuracy | 1 – 0 (correct/ incorrect) |
| 3 | Praxis: Ideation (PrI) | Participant demonstrates all the actions they can think to do in 60 seconds using their bodies, hands, small objects, and a chair | 4 | Tally | Count of novel ideas |
| | | | | Speed | 2 – 0 (high/medium/low) |
| | | | | Variety | 2 – 0 (high/medium/low) |
| | | | | Complexity | 2 – 0 (high/medium/low) |
| 4 | Praxis: Positions (PrP) | Participant imitates static positions demonstrated by the examiner | 24 | Body | 2 – 0 (correct/approximate/incorrect) |
| | | | | Hands | 2 – 0 (correct/approximate/incorrect) |
| | | | | Face | 2 – 0 (correct/approximate/incorrect) |
| 5 | Postural Control (PC) | Participant assumes and maintains a variety of positions and completes reaching tasks | 31[2] | Accuracy | 2 – 0 (correct/approximate/incorrect), time (seconds) |
| | | | | Time (Maintaining Prone Extension and Supine Flexion) | 2: Participant maintains position for 30 seconds<br>1: Participant maintains position for 10-29 seconds<br>0: Participant maintains position for less than 10 seconds |
| 6 | Balance (Bal) | Participant assumes and maintains positions with eyes open and closed (e.g., standing on one foot) | 12 | Balance | Items 1 and 2<br>1: Participant maintains position for 10 seconds or more<br>0: Participant maintains position for less than 10 seconds |
| | | | | | Items 3-12<br>2: Participant maintains position for 10 seconds or more<br>1: Participant maintains position for 5 to 9 seconds<br>0: Participant maintains position for less than 5 seconds |
| 7 | Proprioception: Force (PF) | Crayon Items: Examiner makes a mark with a crayon; Participant makes a mark that matches the intensity of the examiner's mark | 10 | Crayon | 2 – 0 (correct/approximate/incorrect) |
| | | | | Rolling Bottle (Hands) | 2: Second attempt lands within the same segment or 1 segment away from the first attempt<br>1: Second attempt lands 2-4 segments from the first attempt |

| # | Test | Brief Description | Items | Item Types | Scoring |
|---|---|---|---|---|---|
| | | Bottle Items: Participant rolls a bottle of rice to the same segment along a marked yoga mat two times, using one hand, two hands, and one foot | | | 0: Second attempt lands five or more segments from the first attempt |
| | | | | Rolling Bottle (Feet) | 2: Second attempt lands within the same segment or 1 segment away from the first attempt<br>1: Second attempt lands 2-4 segments from the first attempt<br>0: Second attempt lands five or more segments from the first attempt |
| 8a | Ocular Motor (OM) | Participant completes a series of eye movements that demonstrate (1) smooth pursuits, (2) ocular stabilization, (3) quick localization, and (4) ocular praxis | 22 | Pursuits | 2 – 0 (correct/approximate/incorrect) |
| | | | | Stabilization | 2 – 0 (correct/approximate/incorrect) |
| | | | | Localization | 2 – 0 (correct/approximate/incorrect) |
| 8b | Ocular Praxis (PrOc) | Participant imitates eye movements demonstrated by the examiner | 8 | Praxis | 2 – 0 (correct/approximate/incorrect) |
| 9 | Praxis Sequences (PrS) | Participant imitates sequences of positions demonstrated by the examiner | 27 | Body | 2 – 0 (correct/approximate/incorrect) |
| | | | | Hands | 2 – 0 (correct/approximate/incorrect) |
| | | | | Face | 2 – 0 (correct/approximate/incorrect) |
| 10 | Bilateral Integration (BI) | Child performs sequences of activities that use both sides of the body, as demonstrated by the examiner | 15 | Accuracy | 2 – 0 (correct/approximate/incorrect) |
| 11 | Praxis: Following Directions (PrFD) | Participant executes static positions based on verbal instructions | 18 | Body | 2 – 0 (correct/approximate/incorrect) |
| | | | | Hands | 2 – 0 (correct/approximate/incorrect) |
| | | | | Face | 2 – 0 (correct/approximate/incorrect) |
| 12 | Vestibular Nystagmus (VN) | Testing the vestibular nystagmus reflex after clockwise and counterclockwise rotation on a spinning board | 6[3] | Clockwise | Time (seconds) |
| | | | | Counterclockwise | Time (seconds) |
| 13 | Visual Praxis: Construction (VPrC) | Participants arrange objects/small furniture to make a “silly room” based on a visual model (photograph) | 38 | Accuracy | 1 – 0 (correctly positioned/incorrectly positioned) |
| 14 | Proprioception: Joint Position (PJP) | One Hand Items: Examiner places participant’s finger on various spots; participant returns the limb to the same spot | 14 | One Hand | 2: Participant marks with pen 0-2cm from target<br>1: Participant marks with pen 3-5cm from target<br>0: Participant marks with pen 6 or more cm from target |
| | | | | Two Hands | 2: Participant’s hands are positioned 0-1 cm from each other vertically on each side of the door |

| # | Test | Brief Description | Items | Item Types | Scoring |
|---|---|---|---|---|---|
| | | Foot Items: Examiner places participant's toe on various spots; participant returns the limb to the same spot<br><br>Two Hands Items: Participant matches the position of two hands at the same time on either side of an open door | | | 1: Participant's hands are positioned 2-4 cm from each other vertically on each side of the door<br>0: Participant's hands are positioned 5 or more cm from each other vertically on each side of the door |
| | | | | Foot | 2: Participant places toe 0-2cm from target<br>1: Participant places toe 3-5cm from target<br>0: Participant places toe 6 or more cm from target |
| 15 | Auditory Localization (AL) | Clicker Items: Examiner presses a clicker next to various parts of the participant's body; participant identifies where they heard the sound<br><br>Table Items: Examiner taps underneath a table one or two times; participant identifies which quadrant of the table was tapped and how many taps they heard | 20 | Clicker | 1 – 0 (correct/incorrect) |
| | | | | Table | 1 – 0 (correct/incorrect) |
| 16 | Tactile Perception: Localization (TPL) | Examiner touches a spot on the participant's arm or hand with either one or two fingers; participant identifies the spot where they were touched and the number of spots they felt | 20 | Accuracy | 1 – 0 (correct/incorrect) |
| 17 | Tactile Perception: Designs (TPD) | Examiner traces a design on the participant's arm or hand with finger, participant copies the design | 24 | Accuracy | 2 – 0 (correct/approximate/incorrect) |
| 18 | Tactile Perception: Shapes (TPS) | Visual Match Items: Participant matches a shape placed in the hand with a visual model on a field of distractors<br><br>Tactile Match Items: Participant matches a shape or textured tile placed in the hand with an identical shape on a field of distractors | 12 | Visual Match | 1 – 0 (correct/incorrect) |
| | | | | Tactile Match | 1 – 0 (correct/incorrect) |

| # | Test | Brief Description | Items | Item Types | Scoring |
|---|---|---|---|---|---|
| 19 | Tactile Perception: Oral (TPO) | Examiner touches a spot on the participant's arm or hand with either one or two fingers; participant identifies the spot where they were touched and the number of spots they felt | 12 | Accuracy | 1 – 0 (correct/incorrect) |
| 20 | Sensory Reactivity (SR) | Participant responds to stimuli in a variety of modalities, including vestibular, proprioceptive, tactile, auditory, and olfactory[4] | 97 | Auditory | 1 – 0 (normal reactivity/hyper OR hyporeactivity) |
| | | | | Movement/Gravity | 1 – 0 (normal reactivity/hyper OR hyporeactivity) |
| | | | | Tactile | 1 – 0 (normal reactivity/hyper OR hyporeactivity) |
| | | | | Olfactory | 1 – 0 (normal reactivity/hyper OR hyporeactivity) |

[1]Each item scored for accuracy and borders; most items scored for segmentation and jogs
[2]Several postural control items scored based on observations conducted during other tests
[3]To prevent choking/swallowing, objects are fixed to the top of water bottles
[4]17 items scored directly; the remaining 80 items scored based on observations conducted during other tests

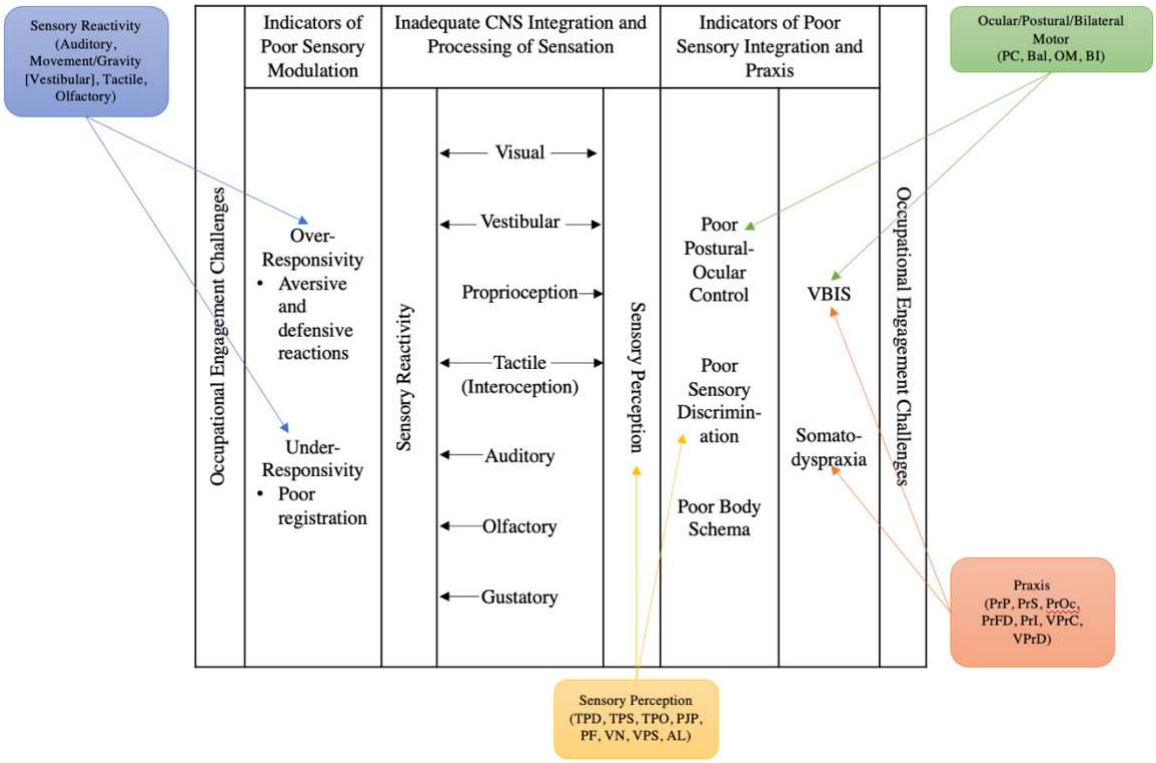



**Figure 1.2**

*EASI Tests Mapped onto the Bundy & Lane (2020) SI Model*

Currently, international normative data collection is underway for the EASI tests in 18 different languages and 81 countries. This suite of instruments may represent a new gold standard for the evaluation of SI functions. Supplementary File 1 contains the full suite of EASI tests.

**Validity and Reliability of EASI: A Need for Further Investigation**

EASI is a promising instrument for the measurement of SI constructs. However, before normative data can be published and used for clinical interpretation, we must evaluate if it is psychometrically sound. Very few studies have examined the validity and reliability of data gathered with this suite of tests. To be clinically useful, assessments must produce valid and reliable data (Brown, 2010). Prinsen and colleagues (2018), authors of the Consensus-based

Standards for the Selection of Health Measurement Instruments (COSMIN), defined validity as "the degree to which an [instrument] measures the construct(s) it purports to measure" (p. 11). Validity is a unified concept: while there are many sources of validity, all sources of validity evidence contribute to the usefulness and appropriateness of a scale for a particular context (American Educational Research Association [AERA], 2014). Construct validity specifically refers to the extent to which data from an instrument is an accurate and meaningful representation of the construct to be measured (Bond, Yan & Heene, 2020; Brown, 2010; Prinsen et al., 2018;). Construct validity is essential for answering a critical validity question: does the instrument measure the construct we intend to measure?

Reliability refers to "the degree to which the measurement is free from measurement error" (Prinsen et al., 2018, p. 11). Data must be reproducible under a variety of circumstances (AERA, 2014). Like validity, reliability can be examined using several methods, including stability over test episodes (i.e., test-retest reliability), stability across raters (i.e., inter-rater reliability), and stability within raters (i.e., intra-rater reliability). Furthermore, the items that comprise a test should be sufficiently related so that any subset of items produces the same measurement as the overall test (i.e., internal reliability).

Mailloux et al. (2018) demonstrated construct validity of EASI data using known-groups analysis. They compared scores for 20 children with known or suspected SI dysfunction with scores for 20 typically developing children. Across all tests except TPS and PF, typically developing children showed significantly better scores than children with SI dysfunction. This provided early evidence for construct validity; however, the tests have undergone significant revisions since this study.

Since the publication of Mailloux et al. (2018), the EASI research group conducted several studies examining the measurement properties of EASI data. The studies comprise data collected with several hundred American children (both typically developing and children with known/suspected sensory integration problems; sample sizes varied across tests). I conducted Rasch analyses on data from each of the EASI tests (Mailloux et al., 2021; Schaaf et al., in press; Grady-Dominguez, unpublished data). As a result of the findings from these analyses, we shortened the EASI tests and restructured scoring procedures. After these revisions, we found strong evidence for construct validity and internal reliability for data from most of the tests. Normative data collection used the revised test versions.

## The Present Study: A Rasch Analysis of EASI

Given the need for additional evidence of validity and reliability of data collected using EASI, my dissertation includes Rasch analyses of data from the EASI tests. The Rasch psychometric model is a latent-trait item response theory (IRT) model that constructs interval-level measures of both item difficulty and person ability from ordinal data (Bond, Yan, & Heene, 2020; Wright & Stone, 1979). The Rasch model is based on three assumptions: (1) that all test items comprise a unidimensional construct that represents a single underlying latent trait; (2) that easier test items are easier for all people, and (3) that people with more ability (i.e., more of the latent trait) are more likely to succeed on harder items. The Rasch model uses a log-odds transformation to generate person and item measures based on the probability that a person with some ability will obtain any given score on an item; these measures are expressed along a common log-odds unit (logit) scale. The Rasch model (1) tests the hypothesis that the items measure a unidimensional construct, (2) produces an ordered set of items and persons, and (3) allows the user to identify items and responses that fail to conform to the Rasch expectations

(Bond, Yan & Heene, 2020; Fox & Jones, 1998).

Additionally, Rasch allows us to examine evidence for *invariance* of measures (Bond, Yan & Heene, 2020). Invariance is a component of construct validity that refers to the stability of item difficulty measures regardless of off-dimensional factors. For items, for example, we can evaluate if item measures remain stable despite the language that the test is presented in, the sex of the child, or the child's IQ. When item measures are not stable, this indicates bias among the test items that may lead to invalid scores. In this study, I evaluated whether sex introduced significant bias into the study: I examined the hypothesis that item measures would be approximately the same for both male and female children.

The use of Rasch measurement theory is appropriate for evaluating performance-based tests and tests, such as EASI, that measure developmental phenomena (Bond, Yan & Heene, 2020). However, in pediatric occupational therapy, most tests have not been rigorously examined using item response theory (IRT) models. Rather, instrument developers often select classical test theory (CTT) instead. Classical test theory assumes that the tester's true score is represented by the sum of their correct responses and some degree of measurement error (Hambleton & Jones, 1993). This theory, on its face, assumes that all items are equally difficult – a risky assumption in performance-based examinations. Moreover, most classical test theory primarily examines total test score. One parameter IRT models such as Rasch analysis, on the other hand, conceptualize test items as a representation of some latent trait that cannot be directly observed, but can be operationalized via test items of varying difficulty. This latent trait is often called the test-taker's "ability". The test taker's score, then, is considered a function of the difference between the difficulty of test items and the test-taker's ability. (Two parameter and three parameter IRT models have been developed; these are outside the scope of this dissertation).

Based on these definitions, it is clear that IRT models such as Rasch have benefits over CTT when examining developmental and/or performance-based latent traits with multiple levels of difficulty.

Rasch analyses have been used to examine measurement properties of other performance-based tests used in pediatric occupational therapy, including Pediatric Balance Scale (PBS; Darr et al., 2015), Peabody Developmental Motor Scales, 2nd edition (PDMS-2; Valenti & Zanella, 2022), and Bruininks-Osteresky Test of Motor Performance, 2nd edition (BOT-2; Wuang et al., 2009). However, these studies are generally small-scale and not conducted by the instrument developers; therefore, they are often not used during the development of tests. My dissertation contributes to the development of EASI by employing Rasch analysis to (1) examine measurement properties and (2) suggest revisions to these instruments.

**Positionality Statement**

As a member of the EASI development team, I am uniquely situated to conduct these analyses. I first became involved with the EASI project in 2018, when I was working as a GRA for Dr. Bundy during my master's thesis. At the time, the EASI development team (led by Dr. Zoe Mailloux, Dr. Susanne Smith-Roley, Dr. Diane Parham, and Dr. Roseann Schaaf) had developed items based on existing SIPT tests as well as other tests that added to the original constructs measured by SIPT (e.g., the Test of Ideational Praxis [TIP]). They previously completed the Mailloux et al. (2018) study (described earlier), otherwise had limited information about the validity and reliability of these tests. I conducted Rasch analyses to (1) identify redundant items for deletion; (2) revise the order of items to follow empirical item difficulties; and (3) establish early evidence for validity and reliability. These analyses resulted in the final tests that were used for normative data analysis.

In addition to my existing involvement with the EASI, I am well-suited for this study because of my experience using Rasch analyses to evaluate evidence for psychometric properties of novel tests. I have published two manuscripts using the Rasch model with two additional pediatric assessments (Grady-Dominguez et al., 2019; Grady-Dominguez et al., 2020). Additionally, I have contributed to unpublished analyses of quite a few other instruments, including: the Test of Playfulness and Test of Environmental Supportiveness (ToP and ToES; Skard & Bundy, 2008), Drive Safe Drive Aware (Kay & Bundy, 2009), the Children's Playfulness Scale (Barnett, 1991) and others.

Perhaps most importantly, I am committed to this work because I believe strongly in the value of assessment as a foundation for clinical practice and research. Without tools that produce valid and reliable data, we cannot adequately answer questions about intervention efficacy and effectiveness. Assessment is foundational to all parts of occupational therapy research and practice; without it, we are essentially blind to the power of our interventions and approaches. My convictions are also influenced by occupational science and rehabilitation science, two different but complementary fields of study that I have been immersed in for the past four and a half years. In Appendix A, I expand in great detail on the influences of occupational science and rehabilitation science on this dissertation. Furthermore, I describe the contributions of this dissertation to each of these fields. Notably, I wrote this discussion before the completion of the dissertation.

I also believe that it is critically important for occupational therapists to be at the center of designing and validating occupational therapy instruments. In many cases, statistical analyses are conducted by statisticians who (while enormously knowledgeable in their own fields) have little understanding of the constructs measured by our assessments. Validity and reliability are

circumstantial and dynamic; we can only establish that the data we collect is useful for the purposes we need. Therefore, occupational therapists must partner closely with statisticians or (better yet) master the techniques necessary to analyze our own assessment tools.

**Research Questions**

Given (1) the need for evidence of the validity and reliability of normative data collected using the EASI (2) and my unique position to complete these analyses, I have applied the Rasch model to answer the following questions:

(1) What is the evidence for construct validity of the data collected using the EASI tests?

   a. Do the test items demonstrate uniformly positive point-measure correlations (i.e., do scores on each item correlate with overall test score?)
   b. Do 95% of items demonstrate adequate fit to the Rasch model?
   c. Do 95% of children demonstrate adequate fit to the Rasch model?
   d. Do the Rasch-generated step thresholds within rating scales progress in an orderly fashion?
   e. Does Rasch principal components analysis of standardized residuals (PCA) reveal meaningful secondary dimensions in the data?
   f. Does differential item functioning (DIF) reveal invariance in item difficulty for children based on sex?
   g. Do the items form a logical hierarchy with sufficient item difficulty variation to match sample ability levels?
   h. Do test-takers form a logical developmental hierarchy (i.e., do scores increase with increasing age?)

(2) What is the evidence for internal reliability of data collected using the EASI tests?

a. Does the data collected using the test demonstrate adequate internal reliability based on the Rasch person reliability index?
b. Does the data collected using the test reliably distinguish at least two levels of sensory integration based on the number of strata associated with the measure?

## Structure of the Dissertation

The dissertation document comprises seven chapters. In this chapter (*Introduction*), I provided a brief overview of SI theory, described the importance of measurement for SI theory development and intervention, and introduced the purpose and research questions that this dissertation will fulfill. In the second chapter (*Literature Review*), I presented the psychometric properties of alternative instruments to the EASI using the COSMIN framework. Based on this review, I concluded that no existing test sufficiently fulfills the need for an SI-specific, internationally normed, clinically useful, valid, and reliable instrument. The third, fourth, and fifth chapters report the results of the Rasch analyses. These chapters are divided into three manuscripts. In Chapter 3 (*Manuscript 1: Praxis*), I evaluated the seven praxis tests of EASI. In Chapter 4 (*Manuscript 2: Sensory Perception*), I evaluated the eight tests that measure aspects of sensory perception. In Chapter 5 (*Manuscript 3: Motor*), I evaluated the four motor tests. In the final chapter (*Conclusion*) I briefly summarized major themes in the findings, described unforeseen problems in the analyses, and suggested future directions for the research. I also expand on how the dissertation process impacted me as a researcher.

Readers may note that, despite careful consideration of sensory reactivity instruments in the literature review, I ultimately omitted the single sensory reactivity EASI test (SR) from this analysis. I made this decision after careful consideration of the SR data; I ultimately decided that

these data did not meet the assumptions of the Rasch model and must be evaluated differently. For further details, please review Chapter 6: *Conclusion*.

# CHAPTER 2: LITERATURE REVIEW

EASI fills a critical gap in sensory integration theory and practice. For over four decades, SIPT served as the only gold standard instrument for clinical measurement in sensory integration (Schaaf et al., 2018). However, advancements in theory and a greater demand for internationally-normed, cost-effective assessments have given rise to EASI – a new suite of instruments that have the potential to improve upon SIPT (Mailloux et al., 2018). This study will establish the validity and reliability of normative data collected using EASI, which will serve as the basis for scoring and interpreting EASI results for children across the globe.

The importance of this study hinges upon the necessity for EASI. In this review of literature, I examine the evidence supporting the validity and reliability of existing tools used to evaluate sensory integration, including SIPT and other instruments used by occupational therapists and related practitioners. Ultimately, this review demonstrates the inadequacies of these instruments and, therefore, the need for EASI.

## Methods

I employed the Consensus-based Standards for the Selection of Health Measurement Instruments (COSMIN; Prinsen et al., 2018) framework to conduct a systematic review of literature examining instruments relevant to sensory integration. Figure 1 shows a schematic summary of the approach I used. The sections that follow detail the steps I took to evaluate the literature.

### Instrument Selection

First, I identified 26 instruments that examine aspects of sensory integration in children. For this review of literature, I included instruments that provided clinical information relevant to

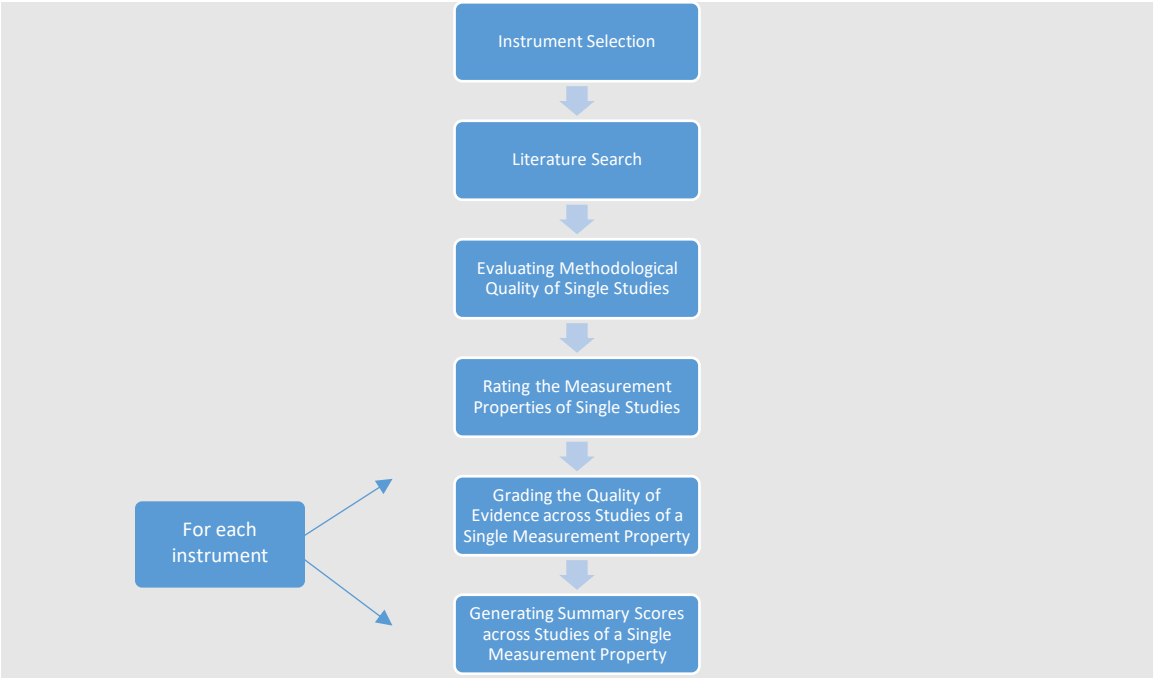



**Figure 2.1**

*Schematic Summary of the Literature Review Approach*

sensory integration constructs (i.e., sensory perception, sensory reactivity, and sensorimotor/praxis abilities). These instruments were not all specific to sensory integration theory; for example, I included measures of gross motor ability because they allow clinicians to observe the influence of sensorimotor abilities (e.g., balance, bilateral integration) on motor skills. I included instruments useful for measurement of children ages 3-12 to reflect the wide range of developmental levels of sensory integration. I excluded instruments (1) that are not appropriate for use with at least one year group between 3-12, (2) that are superseded by a revision/new edition, (3) for which standardized administration procedures are not available either commercially or through direct communication with the authors, and (4) that are designed specifically for a single diagnostic group.

## Literature Search

To conduct this review, I searched the following databases: Google Scholar, CINAHL, PubMed, ERIC, and PsycInfo. Table 1 contains keywords and variants I applied in this search. I also drew upon clinical knowledge of existing measures, and I hand-searched manuals of available instruments.

**Table 2.1**

*Search Terms used in the Literature Review*

| Keyword | Variants |
|---|---|
| | All Searches |
| Assessment | Measure, Measurement, Instrument, Tool, Paradigm |
| Reliability | Reliab*, Reliable, Inter-rater, Intra-rater, Test-retest, Split-half, Cronbach's alpha |
| Validity | Valid*, Construct, Known-groups, Discriminative, Face, Content, Criterion, Responsiveness, Rasch |
| Psychometrics | Psychometric*, Clinimetric* |
| | Modalities |
| Visual | Sight, vision |
| Auditory | Audition, sound, hearing |
| Proprioceptive | Proprioception, body sense, position sense, body awareness |
| Vestibular | Balance, head position, gravitational, gravit* |
| Tactile | Touch, pressure, pain, sensation, temperature |
| Olfactory | Smell |
| Gustatory | Taste |
| | Constructs |
| Perception | Awareness, localization |
| Reactivity | Responsiv*, hyper-reactiv*, hypo-reactiv*, under-reactiv*, over-reactiv*, hyper-responsiv*, hypo-responsiv*, under-responsiv*, over-responsiv* |
| Praxis | Motor planning, coordinat* |
| Sensorimotor | Postural, ocular, bilateral integration, balance, hand-eye, eye-hand, stability, equilibrium |

After compiling a list of relevant instruments, I conducted a literature search for articles evaluating the psychometric properties of these instruments. I included full-text, peer-reviewed articles that evaluated at least one of the following measurement properties: internal consistency,

reliability (including test-retest reliability, inter-rater reliability, intra-rater reliability), content validity, construct validity (including structural validity, hypothesis-testing validity, cross-cultural validity/measurement invariance), or responsiveness. See following sections for more detail regarding these measurement properties. I excluded articles (1) that were not available in English, (2) that did not evaluate children within at least one year group between 3-12 years, (3) only evaluated children with a diagnosis that may obscure sensory reactivity/perception or sensorimotor/praxis ability (e.g., cerebral palsy or blindness).

**Methodological Framework and Taxonomy: COSMIN**

The Consensus-Based Standards for the Selection of Health Measurement Instruments (COSMIN) framework provides theoretical and practical organization for this literature review. COSMIN refers to a set of standards developed by an expert panel to evaluate the content and measurement properties of measurement instruments (Prinsen et al., 2018). Although the COSMIN framework was originally designed to evaluate the measurement properties of patient-reported outcome measures (PROMs; Prinsen et al., 2018), previous researchers have adapted this framework to suit performance-based/observational measures as well (e.g., Griffiths et al., 2018). In this review, I employed the COSMIN framework for both patient/proxy-reported outcome measures (e.g., SPM) and clinician-rated measures (e.g., BOT-2).

The COSMIN framework divides measurement properties into three overarching categories: properties related to reliability, properties related to validity, and properties related to responsiveness. These categories are further divided into types of reliability, validity, and responsiveness.

*Reliability* refers to "the degree to which the measurement is free from measurement error" (Prinsen et al., 2018, pg. 11). Measurement properties included under reliability are

internal consistency, reliability (further divided into test-retest, inter-rater, and intra-rater reliability), and measurement error. Table 2.2 contains definitions and statistical approaches for evaluating each type of reliability.

**Table 2.2**

*Types of Reliability and Statistical Approaches for Evaluating Reliability*

| Measurement Property | Definition | Methods for Evaluating[1] |
|---|---|---|
| Internal Consistency | The degree of inter-relatedness among the items | Cronbach's alpha, KR-20, split-half |
| Reliability | The extent to which scores for test-takers who have not changed are the same for repeated measurements under several conditions (see below) | Correlations (ICC, Pearson's *r*, Spearman's *rho*) |
| *Test-retest* | Reliability over time | |
| *Inter-rater* | Reliability when scored by different people on the same occasion | |
| *Intra-rater* | Reliability when scored by the same person on different occasions | |
| Measurement error | The systematic and random error of a patient's score that is not attributed to true changes in the construct to be measured | Standard error of measurement (SEM), smallest detectable change (SDC), limits of agreement (LoA)[2] |

*Note.* [1]This list is not exhaustive, but represents the main methods observed in this literature review. [2]In order to assess reliability based on measurement error, measurement error value must be compared to an established minimal clinically important difference (MCID).

*Validity* refers to "the degree to which a[n instrument] measures the construct(s) it purports to measure)" (Mokkink et al., 2018, p. 11). In the COSMIN framework, validity is divided into four measurement properties: content validity, structural validity, construct validity, and criterion validity. Construct validity is divided once again into three types: structural validity, hypothesis-testing validity, and cross-cultural validity/measurement invariance. Table 2.3 contains definitions and statistical approaches for evaluating each type of validity.

**Table 2.3**

*Types of Validity and Statistical Approaches for Evaluating Validity*

| Measurement Property | Definition | Methods for Evaluating[1] |
|---|---|---|
| Content Validity | The degree to which the content of an instrument is an adequate reflection of the construct to be measured | See note |
| Construct Validity | The degree to which the scores of an instrument are consistent with hypotheses based on the assumption that the instrument validly measures the construct to be measured | See below |
| *Structural Validity* | The degree to which the scores of an instrument are an adequate reflection of the dimensionality | Factor analysis, Rasch analysis, principal components analysis |
| *Cross-cultural validity/Measurement Invariance* | The degree to which the performance of the items on an instrument are stable when the instrument is translated OR when the instrument is given to different groups that are not expected to differ on the construct to be measured | Differential item functioning (DIF) analysis, multiple-group confirmatory factor analysis (CFA), logistic regression |
| *Hypothesis Testing: Convergent Validity* | The degree to which the scores of an instrument agree with scores of an instrument that measures the same or related constructs | Correlations (Pearson's *r*, Spearman's rho), % agreement, kappa[2] |
| *Hypothesis Testing: Known-groups Validity* | The degree to which the scores of an instrument differ between groups expected to differ on the construct to be measured[1] | Group differences (ANOVA, T-Tests, mean score comparisons)[3] |
| *Hypothesis Testing: Predictive Validity* | The degree to which the scores of an instrument | Regression, correlations |

| Measurement Property | Definition | Methods for Evaluating[1] |
|---|---|---|
| | predict the scores of (1) another instrument, taken at a later time or (2) the same instrument, taken after change is expected to have occurred | |
| *Hypothesis Testing: Validity based on Development* | The degree to which the scores of an instrument reflect changes in the construct that occur with development | Correlations with chronological age/developmental age |

*Note.* The COSMIN approach to evaluating content validity is appropriate for PROMs. However, different approaches are often used to establish content validity for performance-based measures. Unfortunately, these standards are poorly defined (see "A Note about Content Validity." )

[1]This list is not exhaustive, but represents the main methods observed in this literature review

[2]To establish convergent validity, it is less important that relationships are statistically significant, and more important that the magnitude of their relationship meets hypotheses (e.g., evaluators may expect that two instruments have a weak/negative relationship based on existing literature).

[3]Similarly, known-groups validity is best examined by evaluating effect sizes between groups – does the magnitude of the difference between groups agree with the literature about each group?

*Responsiveness* refers to "the ability of a[n instrument] to detect change over time in the construct to be measured" (Mokkink et al., 2018, p. 11). In the COSMIN framework, responsiveness is both the domain and the measurement property. Responsiveness is evaluated by examining the validity of change scores. There are two main approaches appropriate to evaluate responsiveness: criterion and construct; both require longitudinal evaluation of study participants. These approaches are similar to the approaches described for construct and criterion validity (see Table 2.3); however, the cross-sectional scores should be replaced by change scores. Prinsen et al. (2018) describe these approaches in greater detail.

### Evaluating the Methodological Quality of Each Study

The COSMIN framework provided both a taxonomy and a methodological approach for this literature review. After I completed my literature search for articles/manuals evaluating the measurement properties of instruments related to sensory integration, I employed COSMIN's Risk of Bias checklist (Prinsen et al., 2017) to evaluate the methodological quality of each study.

The Risk of Bias checklist contains standards related to study design and statistical approach for each measurement property. See Supplementary File 2 for the COSMIN Risk of Bias Checklist.

I employed the Risk of Bias Checklist to evaluate studies for this literature review. Based on these standards, I rated studies as "Very Good", "Adequate", "Doubtful", or "Inadequate." Notably, each article/manual included in this review may comprise multiple studies (e.g., an article that examines both construct validity and internal consistency reliability).

**Rating Measurement Properties for Each Study**

After rating the methodological quality for each study, I used the COSMIN framework's criteria for adequate evidence for each measurement property. COSMIN uses a three-point rating system: sufficient (+), (-) insufficient, (?) indeterminate. Table 2.4 establishes criteria for each measurement property.

**Table 2.4**

*Criteria for adequate measurement properties, drawn from Prinsen et al. (2018)*

| Measurement Property | | Criteria |
|---|---|---|
| Internal Consistency | + | At least low evidence for sufficient structural validity[1] AND Cronbach's alpha(s) ≥ .70 for each unidimensional scale or subscale |
| | ? | Criteria for "At least low evidence for sufficient structural validity" not met |
| | - | At least low evidence for sufficient structural validity[1] AND Cronbach's alpha(s) < .70 for each unidimensional scale or subscale |
| Reliability | + | Reliability coefficient ≥ .70 |
| | ? | Reliability coefficient not reported |
| | - | Reliability coefficient < .70 |
| Measurement Error | + | Smallest Detectable Difference (SDC), Standard Error of Measurement (SEM) or Limits of Agreement (LoA) < Minimum Clinically Important Difference (MCID)[2] |
| | ? | MCID not defined |
| | - | SDC, SEM or LoA > MCID[2] |

| Measurement Property | | Criteria |
|---|---|---|
| Structural Validity | + | **Confirmatory Factor Analysis (CFA)**:<br>Comparative Fit Index (CFI) or Tucker-Lewis Index (TLI) or comparable measure > 0.95 OR Root Mean Square Error of Approximation (RMSEA) < 0.06 OR Standardized Root Mean Square Residual (SRMR) < 0.082<br>**Item Response Theory (IRT)/Rasch:**<br>No violation of unidimensionality: CFI or TLI or comparable measure > 0.95 OR RMSEA < 0.06 OR SRMR < 0.08<br>AND<br>no violation of local independence: residual correlations among the items after controlling for the dominant factor < 0.20 OR Q3's < 0.37<br>AND<br>no violation of monotonicity: adequate looking graphs OR item scalability > 0.30<br>AND<br>adequate model fit:<br>IRT: $\chi 2$ >0.01<br>Rasch: infit and outfit mean squares ≥ 0.5 and ≤ 1.5 OR Z-standardized values > -2 and <2 |
| | ? | **CFA**: Not all information for '+' reported<br>**IRT/Rasch**: Model fit not reported |
| | - | Criteria for '+' not met |
| Cross-cultural Validity/Measurement Invariance | + | No important differences found between group factors (such as age, sex, language) in multiple group factor analysis OR no important Differential Item Functioning (DIF) for group factors |
| | ? | No multiple group factor analysis OR DIF analysis performed |
| | - | Important differences between group factors OR DIF was found |
| Criterion Validity | + | Correlation with gold standard ≥ 0.70 OR (sensitivity/specificity analysis) Area Under Curve (AUC) ≥ 0.70 |
| | ? | Not all information for '+' reported |
| | - | Correlation with gold standard < 0.70 OR AUC < 0.70 |
| Hypotheses Testing for Construct Validity | + | The result is in accordance with the hypothesis[3] |
| | ? | No hypothesis defined (by the review team) |
| | - | The result is not in accordance with the hypothesis[3] |
| Responsiveness | + | The result is in accordance with the hypothesis[3] OR AUC ≥ 0.70 |

| Measurement Property | | Criteria |
|---|---|---|
| | ? | No hypothesis defined (by the review team) or AUC not defined |
| | - | The result is not in accordance with the hypothesis[3] OR AUC < 0.70 |

*Note*. [1]Internal consistency should only be evaluated for unidimensional scales/subscales. If unidimensionality was not established, I still rated the studies, but included a footnote regarding potential doubtfulness of these results.
[2]MCID may be established in the study under review or in another study.
[3]Hypothesis established by the study authors *or* the reviewer.

### Grading the Quality of Evidence

The COSMIN framework also provides a system for evaluating the quality of the body of evidence for each measurement property (Prinsen et al., 2018; Mokkink et al., 2020). COSMIN employs a four-point quality scale. Table 2.5 contains definitions for this quality scale.

**Table 2.5**

*Definitions of Quality Levels, drawn from Prinsen et al. (2018)*

| Quality Level | Definition |
|---|---|
| High | We are very confident that the true measurement property lies close to that of the estimate of the measurement property. |
| Moderate | We are moderately confident in the measurement property estimate: the true measurement property is likely to be close to the estimate of the measurement property, but there is a possibility that it is substantially different. |
| Low | Our confidence in the measurement property estimate is limited: the true measurement property may be substantially different from the estimate of the measurement property. |
| Very Low | We have very little confidence in the measurement property estimate: the true measurement property is likely to be substantially different from the estimate of the measurement property. |

I graded the body of evidence for each measurement property (for each instrument) based on three factors: risk of bias, (in)consistency, (im)precision, and (in)directness. To grade the evidence, I started with a rating of "High". Then, I downgraded the results based on the criteria described in Table 2.6. Of note, the COSMIN framework also recommends grading evidence

based on indirectness: the relevance of the study population to the population of interest in the review. However, because I excluded studies that were not relevant to the population of interest, I did not grade the evidence using this criterion.

**Table 2.6**

*COSMIN Approach to Grading Quality of Evidence, drawn from Prinsen et al. (2018)*

| Grading Factor | Criteria for Downgrading[1] |
|---|---|
| Risk of Bias | -0: There are multiple studies of at least adequate quality, or there is one study of very good quality available<br>-1: There are multiple studies of doubtful quality available, or there is only one study of adequate quality<br>-2: There are multiple studies of inadequate quality, or there is only one study of doubtful quality available<br>-3: There is only one study of inadequate quality available |
| Inconsistency | -1: The results across studies are slightly inconsistent and no explanation for inconsistency is evident<br>-2: The results across studies are very inconsistent and no explanation for inconsistency is evident |
| Imprecision | -1: Total sample (across studies) = 50-100<br>-2: Total sample (across studies) < 50 |

*Note*. [1]The numbers in this column refer to levels of evidence (i.e., a score of -1 would lower the grade from High to Moderate).

### Generating Summary Scores

After rating the overall quality of the body of evidence, I determined a summary score for each category. The COSMIN framework provides little guidance for quantitatively pooling results across studies. Therefore, for the present study, I determined the summary score by evaluating the adequacy of evidence across studies. In other words, if all studies examining a particular measurement property showed sufficient (+) evidence for this property, I rated the evidence for this property as (+). If all studies showed insufficient (-) evidence, I gave a rating of (-). If some studies showed sufficient (+) evidence and some showed insufficient (-) evidence, I gave a rating of inconsistent (+/-). In a few cases, when a large number of studies were sufficient

(+) while only one or two studies of doubtful/inadequate methodological quality were insufficient (-), I gave a rating of (+) but provided a footnote identifying the anomalous studies.

### Evaluating Content Validity

Content validity refers to "the degree to which the content of an instrument is an adequate reflection of the construct to be measured" (Mokkink et al., 2018, p. 11). The COSMIN framework establishes test content as the most important source of validity evidence (Prinsen et al., 2018). In order to provide a valid measurement of the construct of interest, an instrument must be comprised of items that represent that construct. However, the methods for evaluating content validity described in the COSMIN manual are limited to PROMs and, therefore, not appropriate for many of the instruments in this literature review.

Evaluation of content validity is often a subjective judgment and requires examination of several sources of evidence (Portney, 2020). The content validity of an instrument may be strengthened when the authors develop items based on a strong theoretical model and/or a thorough review of literature. Often, expert and/or user review panels provide evidence for content validity. Furthermore, strategies to revise and refine the item set may support content validity. Factor analyses, Rasch analyses, readability analyses, and differential item functioning, for example, may be sources of content validity when the authors use these approaches to remove, re-order, or add items. Of note, these analytical approaches to selecting a set of items should always be integrated with a theoretical knowledge of the construct to be measured.

For each instrument included in this review, I provide an overview of evidence for content validity. Table 2.7 contains each source of evidence I considered.

## Structure of this Review

This literature review is organized into four parts. In the first section, I review and

**Table 2.7**

*Sources of Evidence for Construct Validity*

| Non-Analytical Sources of Evidence |
|---|
| Items derived from theoretical model |
| Items derived from review of literature |
| Items derived/revised based on feedback from potential users |
| Items derived/revised based on feedback from expert panel |
| Pilot study conducted to revise/refine items/instructions |
| **Analytical Sources of Evidence** |
| Items revised based on conventional item analyses (e.g., item discrimination, item difficulty, floor/ceiling effects, item reliability)[1] |
| Items revised based on item response theory or Rasch analyses[1] |
| Items revised based on factor analyses[1] |
| Items revised based on readability analyses |
| Items revised based on differential item functioning analyses[1] |

[1]These approaches may be used to establish other measurement properties (e.g., reliability or structural validity); they are considered sources of evidence for content validity when the instrument authors use the results of these analyses to revise/refine items/instructions

critique evidence for validity and reliability of the Sensory Integration and Praxis Tests, the current gold standard instruments for evaluating SI functions. In the second, third, and fourth sections, I review and critique evidence for validity and reliability of other instruments useful for evaluating three categories of SI function: (1) sensorimotor/praxis functions, (2) sensory reactivity, and (3) sensory perception. Appendix B contains specific results from each of the studies evaluated as a part of this review. The following sections summarize the compiled results of these studies.

## Sensory Integration and Praxis Tests

SIPT (Ayres, 2005) are a suite of 17 individual tests designed to evaluate sensory integration in children ages 4 to 8 years. SIPT were designed by Ayres over three decades, beginning in the 1960s with a group of instruments called the Southern California Sensory Integration Tests (SCSIT). Currently, these tests represent the most thorough and theory-driven option for evaluation of children's sensory integrative functions. Table 2.8 describes the tests that comprise SIPT.

**Table 2.8**

*SIPT Descriptions*

| SIPT Test | Area of SI Function | Description |
|---|---|---|
| Space Visualization (SV) | Sensory Perception (Visual) | Evaluates visuospatial relationships (i.e., mental manipulations of objects in space) by asking the child to identify which of two blocks will fit into the hollow on a form board |
| Figure-Ground Perception (FG) | Sensory Perception (Visual) | Evaluates figure-ground perception (i.e., ability to identify a figure on a rival background) by asking the child to identify embedded, line-drawn figures within a larger line-drawn background |
| Manual Form Perception (MFP) | Sensory Perception (Tactile/Visual) | Evaluates form perception and stereognosis in two parts; first, the child identifies the visual counterpart to a hidden three-dimensional form in the child's hand; second, the child matches a hidden form in one hand with an identical form in the other hand |
| Kinesthesia (KIN) | Sensory Perception (Proprioception) | Evaluates the child's proprioceptive perception by moving the child's limb in space, returning the limb to neutral, and asking the child to match the position |
| Finger Identification (FI) | Sensory Perception (Tactile) | Evaluates tactile perception by touching the child's fingers in one or two places, vision occluded and asking the child to touch the same spot |
| Graphesthesia (GRA) | Sensory Perception (Tactile) | Evaluates graphesthesia by asking the child to duplicate a design traced on the back of the child's hand with the examiner's finger |
| Localization of Tactile Stimuli (LTS) | Sensory Perception (Tactile) | Evaluates tactile perception based on the child's ability to identify the location of a spot on their hand/arm that was previously touched by the examiner |
| Postrotary Nystagmus (PRN) | Sensory Perception (Vestibular) | Evaluates the vestibular-ocular relationship by recording the duration of postrotational nystagmus after spinning the child on a spinning board |
| Standing and Walking Balance (SWB) | Sensorimotor (General Motor) | Evaluates balance based on the child's ability to copy a variety of positions or complete movements with eyes open and closed |
| Motor Accuracy (MAc) | Sensorimotor (Visual Motor) | Evaluates eye-hand coordination based on the child's ability to draw along a dotted line |
| Design Copying (DC) | Sensorimotor (Visual Motor, Praxis) | Evaluates visual motor control based on the child's ability to copy a design with and without a dot grid |
| Postural Praxis (PPr) | Sensorimotor (Praxis) | Evaluates praxis based on visual imitation based on the child's ability to assume postures demonstrated by the examiner |
| Praxis on Verbal Command (PrVC) | Sensorimotor (Praxis) | Evaluates praxis on verbal command based on the child's ability to assume positions based on verbal requests |

| SIPT Test | Area of SI Function | Description |
|---|---|---|
| Constructional Praxis (CPr) | Sensorimotor (Praxis) | Evaluates ability to assemble objects based on a visual model by asking the child to duplicate (1) block structures built by the examiner and (2) a pre-assembled structure |
| Sequencing Praxis (SPr) | Sensorimotor (Praxis) | Evaluates ability to execute a series of planned movements based on the child's ability to imitate a series of hand or finger movements demonstrated by the examiner |
| Oral Praxis (PrOc) | Sensorimotor (Praxis) | Evaluates praxis based on imitation based on the child's ability to assume positions of the tongue, lips, cheeks, or jaw as demonstrated by the examiner |
| Bilateral Motor Coordination (BMC) | Sensorimotor (Praxis and general motor function) | Evaluates ability to smoothly coordinate both sides of the body based on the child's ability to imitate novel motor sequences that require both hands and/or both feet |

**Reliability of SIPT**

Evidence for reliability of SIPT data is limited. Ayres (2005) conducted test-retest and inter-rater reliability studies for each test; Table 2.9 summarizes these results. While the studies had adequate methodological quality based on the COSMIN framework, small sample sizes decreased the quality of the body of evidence to low or very low.

Nine of the 17 SIPT demonstrated sufficient evidence for test-retest reliability (FI, GRA, SWB, MAc, PPr, PrVC, SPr, PrOc, and BMC). The remaining tests had test-retest reliability coefficients <.70, although two tests approached this value (MFP and CPr). These results suggest that, for at least six SIPT, children may demonstrate markedly different responses to the same items across two test sessions; however, low sample sizes call these results into question. All SIPT demonstrated sufficient evidence for inter-rater reliability when two raters observed the same test session. Although the methodological qualities of these studies were low or very low due to small sample size, the results were generally positive. No studies examined internal consistency, intra-rater reliability, or measurement error of SIPT.

**Table 2.9**

*Reliability of SIPT*

| Assessment | Test-Retest Reliability | | | Inter-Rater Reliability | | |
|---|---|---|---|---|---|---|
| | # | QoE | Rating | # | QoE | Rating |
| SV | 1 | VL | - | 1 | L | + |
| FG | 1 | VL | - | 1 | L | + |
| MFP | 1 | VL | -[1] | 1 | VL | + |
| KIN | 1 | VL | - | 1 | L | + |
| FI | 1 | VL | + | 1 | L | + |
| GRA | 1 | VL | + | 1 | L | + |
| LTS | 1 | VL | - | 1 | L | + |
| PRN | 1 | VL | - | 1 | L | + |
| SWB | 1 | VL | + | 1 | L | + |
| MAc | 1 | VL | + | 1 | L | + |
| DC | 1 | VL | - | 1 | L | + |
| PPr | 1 | VL | + | 1 | L | + |
| PrVC | 1 | VL | + | 1 | L | + |
| CPr | 1 | L | -[1] | 1 | L | + |
| SPr | 1 | VL | + | 1 | L | + |
| PrOc | 1 | VL | + | 1 | L | + |
| BMC | 1 | VL | + | 1 | VL | + |

NOTE. # = number of studies, QoE = quality of evidence across studies, H = high, M = moderate, L = low, VL = very low, + = acceptable, +/- = inconsistent, - = unacceptable
[1]Correlations very close to .70 (r = .69 for MFP and r = .67 for CPr)

## Validity of SIPT

Evidence for validity of the SIPT is relatively strong. SIPT have evidence for content validity, construct validity (i.e., hypothesis-testing validity, structural validity, and cross-cultural validity/measurement invariance), and criterion validity. Of note, most studies examining SIPT are published in the manual; additional studies conducted by reviewers not invested in the instrument would provide further evidence for validity (Greenslade & Coggins, 2016). Furthermore, multiple sources of validity evidence had low methodological quality based on COSMIN criteria; therefore, these results should be interpreted cautiously. Evidence for one test, PRN, is particularly weak. This may be the result of the test scoring; both high and low scores are considered problematic. Absolute value transformations may clarify validity of this in future studies.

### Content Validity

SIPT demonstrate strong content validity. SIPT were based upon a robust foundation in theory, literature, and previous assessments. SIPT is based on Ayres' SI theory (Ayres, 1979; Ayres, 2005), a commonly used and widely studied practice and assessment framework among occupational therapists. The author conducted extensive reviews of literature to select items that adequately represented SIPT constructs. Furthermore, SIPT represents a modification of a previous instrument, SCSIT; Ayres incorporated feedback from examiners to modify this instrument. A team of expert reviewers supported item selection. Overall, these sources of evidence support strong content validity.

### Construct Validity

Evidence for construct validity of SIPT stems from hypothesis-testing studies and structural validity studies. I identified three types of hypothesis-testing validity: convergent validity, known-groups validity, and validity based on age/development. Table 2.10 contains results of the hypothesis-testing studies. The results suggest strong evidence for convergent validity with expected subtests on the Kaufman Assessment Battery for Children (K-ABC). For two tests examining constructional praxis (DC and CPr), Cermak and Murray (1991) demonstrated evidence for convergent validity with four additional tests (Wechsler Intelligence Scale for Children-Revised [WISC-R], Beery-Buktenica Visual Motor Integration Test-Revised [VMI-R], Primary Visual Motor Test, and Rey Osterrieth Complex Figure Test). However, these results were only evidence for a sample of children with learning disabilities; the authors found no relationship between K-ABC and CPr for typically developing children. Given the small sample of this study ($n$ = 39), these results should be interpreted cautiously. A larger sample may

have yielded a more noticeable relationship between SIPT and these tests for typically developing children.

**Table 2.10**

*Construct Validity of SIPT*

| Assessment | Hypothesis Testing: Convergent Validity | | | Hypothesis Testing: Known-Groups Validity | | | Hypothesis Testing: Validity based on Development | | | Structural Validity | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | # | QoE | Rating | # | QoE | Rating | # | QoE | Rating | # | QoE | Rating |
| SV | 1 | L | +[1] | 1 | M | +[3] | 1 | L | + | | | |
| FG | 1 | L | +[1] | 1 | M | +[3] | 1 | L | + | | | |
| MFP | 1 | L | +[1] | 1 | M | +[3] | 1 | L | + | | | |
| KIN | 1 | L | +[1] | 1 | M | +[3] | 1 | L | + | | | |
| FI | 1 | L | +[1] | 1 | M | +[3] | 1 | L | + | | | |
| GRA | 1 | L | +[1] | 1 | M | +[3] | 1 | L | + | 1 | H | + |
| LTS | 1 | L | +[1] | 1 | M | +[3] | 1 | L | + | | | |
| PRN | 1 | L | +[1] | 1 | M | -[3] | 1 | L | ? | | | |
| SWB | 1 | L | +[1] | 1 | M | +[3] | 1 | L | + | | | |
| MAc | 1 | L | +[1] | 1 | M | +[3] | 1 | L | + | | | |
| DC | 2 | M | +/-[2] | 2 | H | +[3] | 1 | L | + | | | |
| PPr | 1 | L | +[1] | 1 | M | +[3] | 1 | L | + | 1 | H | + |
| PrVC | 1 | L | +[1] | 1 | M | +[3] | 1 | L | + | | | |
| CPr | 2 | M | +/-[2] | 2 | H | +[3] | 1 | L | + | | | |
| SPr | 1 | L | +[1] | 1 | M | +[3] | 1 | L | + | 1 | H | + |
| PrOc | 1 | L | +[1] | 1 | M | +[3] | 1 | L | + | 1 | H | + |
| BMC | 1 | L | +[1] | 1 | M | +[3] | 1 | L | + | 1 | H | + |
| Combined Tests | | | | | | | | | | 6 | H | +/- |

NOTE. # = number of studies, QoE = quality of evidence across studies, H = high, M = moderate, L = low, VL = very low, + = acceptable, +/- = inconsistent, - = unacceptable

[1]Comparison instruments = Kaufman Assessment Battery for Children (K-ABC)

[2]Comparison instruments = K-ABC (+), Weschler Intelligence Scales for Children-Revised (+/-), Beery Visual Motor Integration Test-Revised (+/-), Primary Visual Motor Test (+/-), and Rey Osterrieth Complex Figure Test (+/-)

[3]Comparison groups = Autism Spectrum Disorder (ASD), Learning Disabilities (LD), Acquired Brain Injury (ABI), Sensory Integration Dysfunction (SID), Intellectual Disability (ID), Spina Bifida, Reading Disorder, Learning Disorder, Cerebral Palsy (CP) (Note. children with spina bifida did not complete PRN)

Known-groups validity studies demonstrated expected patterns of function and dysfunction for children with a variety of diagnoses (see Table 2.10, footnotes). Generally, the results of these studies matched hypotheses, supporting validity of SIPT. On PRN, only one group (children with ID) demonstrated a marked difference from the normative group. Of note, the sample sizes across clinical groups varied drastically (e.g., seven children with ASD and 195

children with learning disabilities). Therefore, the results for these smaller studies should be considered preliminary.

The final source of hypothesis-testing validity comes from age-based analyses. In the SIPT manual, tables display raw means and smoothed means calculated based on polynomial regression. For both males and females, all but one test demonstrated expected age progressions, with near monotonicity in mean scores across age groups. PRN did not show developmental trends; it is not clear if this was in line with the authors' hypotheses; therefore, the developmental validity of PRN remains questionable.

Six studies have examined the structural validity of SIPT based on factor analyses combining scores across tests. While the methodological quality of each study is adequate or very good, the results from each study suggest slightly different patterns of SI function/dysfunction, thus earning a score of inconsistent (+/-). Ayres (2005) emphasized that the tests should be interpreted as a whole; therefore, understanding the structure of the instruments is critical to establish construct validity. Notably, while the studies do reveal different patterns, some authors found similarities across tests (see Appendix B for further details). Only one study (Lai et al., 1996) examined the internal structure of individual tests (GRA, PPr, SPr, PrOc, and BMC). Using Rasch analysis, the authors found strong evidence for unidimensionality of these tests. Furthermore, they found evidence that, taken together, these five tests may represent a unidimensional praxis construct.

**Criterion Validity**

One study, published in the SIPT manual (Ayres, 2005), evaluated the ability of SIPT to distinguish among typically developing children, children with learning disabilities, and children

with SI dysfunction. The study had adequate methodological quality and an acceptable sample size ($N$ = 293); therefore, there is moderate evidence for adequate criterion validity of SIPT.

**Responsiveness of SIPT**

Only one study (Kimball, 1990) examined the responsiveness of SIPT scores to change. Table 2.11 describes the results of this study, which suggest adequate response to change for only four SIPT: LTS, SWB, PPr, and SPr. Unfortunately, this study only conducted pre-test/post-test group comparisons; this is not an accepted method for evaluating this property based on COSMIN standards. Rather, COSMIN recommends that changes be compared with MCID. Therefore, the SIPT should not be used as an outcome measure. The manual for this instrument emphasizes the use of SIPT as a diagnostic tool rather than an outcome measure – therefore, the limited evidence for responsiveness is less problematic.

**Table 2.11**

*Responsiveness of SIPT*

| Assessment | # | QoE | Rating |
|---|---|---|---|
| SV | 1 | VL | - |
| FG | 1 | VL | - |
| MFP | 1 | VL | - |
| KIN | 1 | VL | - |
| FI | 1 | VL | - |
| GRA | 1 | VL | - |
| LTS | 1 | VL | + |
| PRN | 1 | VL | - |
| SWB | 1 | VL | + |
| MAc | 1 | VL | - |
| DC | 1 | VL | - |
| PPr | 1 | VL | + |
| PrVC | 1 | VL | - |
| CPr | 1 | VL | - |
| SPr | 1 | VL | + |
| PrOc | 1 | VL | - |
| BMC | I | VL | - |

**SIPT: A Critical Review**

While SIPT remains the current gold-standard for evaluation in sensory integration functions, the suite of instruments has several notable limitations. First, evidence for reliability is poor. While the few studies conducted suggest good evidence for most of the tests, these studies are too small to be satisfactory. Reliable data are fundamental to adequate measurement; this presents a major limitation to SIPT.

Evidence for validity of SIPT is generally stronger than for reliability; however, this body of literature is not without fault. While SIPT have a remarkable number of large-scale studies evaluating structural validity, each of these studies draws slightly different conclusions about the structure of SIPT. Only one study examines the internal structure of individual SIPT (Lai et al., 1996). Although SIPT scores are not combined across tests to form composites, the manual emphasizes the importance of interpreting results as a whole – therefore, evidence for structural validity is critical.

Of note, Ayres (2005) provided several sources of validity evidence not considered in the COSMIN framework. First, Ayres conducted an extensive cluster analysis, grouping children based on their SIPT score profiles. Cluster analysis may be considered a form of structural validity, although it is not an accepted method in Prinsen et al. (2018). In these studies, the clusters demonstrated still another pattern compared to the factor analyses (although with some notable overlap); therefore, evidence for structural validity remains questionable. Additionally, a main source of validity claims in the manual stems from inter-test correlations. While inter-test correlations are not considered a source of validity evidence for the COSMIN framework, these relationships may help support the internal structural validity of the SIPT (Prinsen et al., 2018).

Finally, only one low-quality study examined the responsiveness of SIPT (Kimball, 1990). As a result, this measurement property cannot be established for SIPT. Without evidence for responsiveness to change, SIPT should not be used as an outcome measure.

In addition to concerns with the measurement properties of SIPT, the tests are limited in several other critical ways. First, the normative data for SIPT are outdated, and their generalizability is limited. These data, collected exclusively in the United States (US) in the 1980s, serve as the basis for all scoring, reliability, and validity data. Moreover, SIPT are expensive to learn, administer, and score. To administer SIPT, therapists must hold a costly certification in Ayres Sensory Integration (ASI). The publisher holds a proprietary scoring program that costs approximately $37 for each output. Given the expense, SIPT are often inaccessible to therapists, especially in underdeveloped or impoverished communities. Finally, SIPT have been normed only on children ages 4-8 years; scores for younger or older children may have limited validity and/or reliability.

Because of the limitations of SIPT, therapists may be disinclined or unable to use this suite of instruments (Szklut, 2010; Mailloux et al., 2018). As a result, therapists must rely on parent/caregiver reports and clinical observations of children's sensory integrative functions. While these are undeniably important components of thorough analysis, they should not supplant standardized, quantitative measurement (Schaaf & Lane, 2015). Given the concerns of SIPT coupled with the benefits of standardized assessment, therapists sometimes seek alternative instruments to evaluate sensory integration functions.

In the sections that follow, I describe available alternatives to the SIPT. In order to be clinically useful, these assessments must have evidence for validity and reliability; therefore, this review focuses on available literature regarding the measurement properties of these instruments.

I divided this review into three sections: sensorimotor/praxis instruments, sensory reactivity instruments, and sensory perception instruments. I further divided the sensorimotor/praxis instruments by major groupings: general (gross/fine) motor instruments, visual motor instruments, and instruments designed to evaluate specific praxis abilities. Some instruments could be categorized into multiple groupings (e.g., PDMS-2, which contains both general motor and visual motor instruments). Where possible, I included these instruments with their most logical groupings. One instrument, NEPSY-II, could not be well-categorized into any of the sections of this review; the skills evaluated through this test are very diverse, and not all apply to evaluation of sensory integration functions. Furthermore, these tests are not designed to be interpreted as a whole; rather, certain batteries are appropriate for specific groups of children. Therefore, NEPSY-II tests are divided into their appropriate categories.

## Sensorimotor/Praxis Instruments

In Tables 2.12, 19, and 24, I describe 17 instruments relevant to evaluation of sensorimotor/praxis abilities. I divided these into three groups: ten general motor instruments that evaluate elements of gross and fine motor ability, four instruments that specifically evaluate visual motor skills, and four instruments that evaluate praxis (one of these instruments also evaluates visual motor skills). Many of the general motor instruments may be used to observe motor planning/praxis abilities (e.g., instruments with throwing/catching items); however, for simplicity, I categorized these only with the general motor group.

### General Motor Instruments

Twelve instruments examine general motor skills, including gross motor skills (all 12 instruments) and fine motor skills (six instruments): Bruininks-Osteresky Test of Motor Performance – 2nd Edition (BOT-2), Movement Assessment Battery for Children – 2nd Edition

(MABC-2), Peabody Developmental Test of Motor Skills – 2nd Edition (PDMS-2), Test of Gross Motor Development – 3rd Edition (TGMD-3), Miller Function and Participation Scales (M-FUN), De-Gangi Berk Test of Sensory Integration (TSI), Clinical Observations of Motor and Postural Skills – 2nd Edition (COMPS-2), Sensory Integration Clinical Observations (SICO), Quick Neurological Screening Test – 3rd Edition, Revised (QNST-3R), and Pediatric Balance Scale (PBS). Table 2.12 provides details about these instruments. The majority of instruments (11 of 12) are norm-referenced. The PDMS-2 allows for assessment of the youngest children, beginning at birth, while QNST-3R allows assessment of children, adults, and older adults. None of these instruments requires advanced training; instead, users are encouraged to read and review the manual for administration, scoring, and interpretation guidelines.

**Reliability of General Motor Instruments**

Thirty-six studies evaluated the reliability of the 12 general motor instruments. These studies evaluated each type of reliability described in the COSMIN manual: internal consistency, test-retest reliability, inter-rater reliability, and intra-rater reliability. Table 2.13 contains the summary results for test reliability. Most instruments had at least low evidence for internal consistency and test-retest reliability; studies of inter- or intra-rater reliability were less common. In general, BOT-2, MABC-2, and TGMD-3 had the strongest evidence for reliability: high quality evidence suggested strong measurement properties for at least three types of reliability. For each of the remaining instruments, additional high-quality studies should clarify reliability. Eight tests had sufficient evidence of internal consistency in at least one (and often more) adequate or very good study (BOT-2, PDMS-2, TGMD-3, M-FUN, COMPS-2, SOSI-M, QNST-3R, PBS). MABC-2 demonstrated inconsistent internal consistency across studies, with three studies suggesting insufficient internal consistency (Ellinoudis et al., 2011; Hirata et al., 2018;

**Table 2.12**

*Description of General Motor Instruments*

| Assessment | Description | Subscales Assessed | Age Range (years) | Administration Time (minutes) | Type | Training | Scores | Languages |
|---|---|---|---|---|---|---|---|---|
| Bruininks-Osteresky Test of Motor Performance – 2nd Edition (BOT-2; Bruininks & Bruininks, 2005) | Developmental assessment designed to evaluate gross and fine motor skills | - Fine manual control<br>- Manual coordination<br>- Body coordination<br>- Strength and agility | 4-21 | 40-60 | Norm-referenced, performance-based | Manual review | Raw score, scaled score, standard score, percentile rank, age equivalent | English |
| Movement Assessment Battery – 2nd Edition (MABC-2; Henderson et al., 2007) | Assessment designed to identify children with impairments in motor functions | - Manual dexterity<br>- Aiming and catching<br>- Balance | 3-16 | 20-40 | Norm-referenced, performance-based | Manual review | Raw score, standard score, percentile rank, functional category | English |
| MABC-2 Checklist (Henderson et al., 2007) | Proxy-report survey designed to evaluate children's motor skills in everyday contexts; supplement to MABC-2 | - Movement in a static and/or predictable environment<br>- Movement in a dynamic and/or unpredictable environment | 5-11 | 10 | Norm-referenced, caregiver or teacher report | Manual review | Raw score, percentile rank | English |
| Peabody Developmental Motor Skills – 2nd Edition (PDMS-2; Folio & Fewell, 2000) | Developmental assessment designed to assess fine, gross, and visual motor skills | - Reflexes<br>- Stationary<br>- Locomotion<br>- Object manipulation<br>- Grasping<br>- Visual-motor integration | 0-6 | 60 | Norm-referenced, performance-based | Manual review | Raw score, standard score, percentile rank, age equivalent | English |
| Test of Gross Motor Development – 3rd Edition (TGMD-3; Ulrich, 2019) | Developmental assessment designed to evaluate gross motor skills and identify children with gross motor impairments | - Locomotor<br>- Ball skills | 3-10 | 30-40 | Norm-referenced, performance-based | Manual review | Raw score, scaled score, percentile rank, age equivalent | English |

| Assessment | Description | Subscales Assessed | Age Range (years) | Administration Time (minutes) | Type | Training | Scores | Languages |
|---|---|---|---|---|---|---|---|---|
| Miller Function and Participation Scales (M-FUN; Gross/Fine Motor subtests: Go Fishing, Penny Bank, Origami, Snack Time, Statue, Throw and Catch, Ball Balance, Bouncing Ball, Soccer, Jumping; Miller, 2006) | Assessment designed to evaluate motor skills and other skills related to school function and participation | - Gross Motor<br>- Fine Motor | 2.5-7 | 60 | Norm-referenced, performance-based | Manual review | Raw score, scaled score, percentile rank, age equivalent | English |
| DeGangi-Berk Test of Sensory Motor Integration (TSI) | Assessment designed to identify children with sensory integrative dysfunction characterized by subtle motor difficulties | - Postural control<br>- Bilateral motor coordination<br>- Reflex integration | 3-5 | 30 | Criterion-referenced, performance-based | Manual review | Raw score | English |
| Clinical Observations of Motor and Postural Skills (COMPS-2; Wilson et al., 2000) | Standardized clinical observations of motor and postural skills based on Ayres' sensory integration theory | None - assesses six motor/postural components | 5-15 | 15 | Criterion-referenced, performance-based | Manual review | Raw score, age-weighted score | English |
| Sensory Integration Clinical Observations (SICO; May-Benson & Teasdale, 2021)* | Standardized clinical observations of motor and postural skills Ayres' sensory integration theory | None- assesses twenty motor/postural items | 4-12 | 30-45 | Criterion-referenced, performance-based | Advanced training | Raw score | English |
| Structured Observations of Sensory Integration – Motor (SOSI-M; Blanche, Gustavo & Reinoso, 2021) | Standardized clinical observations of motor, postural and ocular skills based on Ayres' sensory integration theory | None- assesses thirty-two motor/postural/ ocular items | 5-14 | 20-40 | Norm-referenced, criterion-referenced, performance-based | Manual review | Raw score, standard score, percentile rank | English, Spanish |
| Quick Neurological Screening Test – 3rd Edition, Revised (QNST-3R; Mutti et al., 2017) | Standardized clinical observations of motor and postural skills designed to identify people with sensorimotor deficits | None - assesses 13 "neurological soft signs" (i.e., tasks with motor/sensory | 5-80+ | 20 | Norm-referenced, criterion-referenced, performance-based | Manual review | Raw score, scaled score, percentile rank | English |

| Assessment | Description | Subscales Assessed | Age Range (years) | Administration Time (minutes) | Type | Training | Scores | Languages |
|---|---|---|---|---|---|---|---|---|
| | | integration elements) | | | | | | |
| Pediatric Balance Scale (PBS; Franjoine et al., 2003) | Assessment designed to evaluate balance skills; a pediatric adaptation of the Berg Balance Scale (Berg et al., 1992) | None - assesses functional balance in the context of everyday tasks | 2-7 | 10-20 | Criterion-referenced, performance-based | Article Review | Raw score | English |

* = not yet published

Hua, Gu, Meng, & Wu, 2013) and four studies suggesting sufficient internal consistency of subscales (Hirata et al., 2018; Valentini, Ramalho, & Oliveira, 2014; Wuang, Su, & Huang, 2012; Wuang, Su, & Su, 2012). The subscales of MABC-2 have relatively few items; this may explain the low internal consistency coefficients in some studies (Tavakol & Dennick, 2011). MABC-2 checklist demonstrated sufficient evidence for internal consistency, but the single study evaluating this measurement property was of inadequate methodological quality; therefore, this measurement property should be interpreted cautiously. Furthermore, five instruments had high internal consistency coefficients, but little evidence for unidimensionality of scales/subscales (MABC-2, M-FUN, COMPS-2, SOSI-M, QNST-3R). Because unidimensionality is a prerequisite for internal consistency, these coefficients should also be interpreted cautiously (Prinsen et al., 2018). At the time of this review, I did not identify any studies evaluating the internal consistency of TSI or SICO.

Eleven instruments demonstrated sufficient evidence for test-retest reliability (BOT-2, MABC-2, PDMS-2, TGMD-3, M-FUN, TSI, COMPS-2, SICO, SOSI-M, QNST-3R, PBS). However, seven instruments had only low/very low-quality evidence for test-retest reliability (M-FUN, TSI, COMPS-2, SICO, SOSI-M, QNST-3R, PBS). These instruments each had only a single study of doubtful or insufficient quality. Most often, I downgraded the methodological quality because of questionable statistical approaches; specifically, authors used Pearson's *r* or Spearman's *rho c*orrelations (rather than ICCs) without sufficient evidence that no change occurred within the study population between test and retest (Prinsen et al., 2018). No studies evaluated the test-retest reliability of the MABC-2 checklist. Although the manual cites adequate test-retest reliability for this instrument, the authors base this conclusion off a previous edition of the Checklist; therefore, I cannot assume these results apply to the current, revised version.

Ten instruments demonstrated sufficient evidence for inter-rater reliability (BOT-2, MABC-2, MABC-2 Checklist, TGMD-3, M-FUN, TSI, COMPS-2, SICO, SOSI-M, PBS). However, as with test-retest reliability, the methodological quality of the body of evidence for many of these tests was low (all except MABC-2 and TGMD-3). Most studies had small sample sizes (e.g., SICO, SOSI-M, TSI) or too few studies (e.g., COMPS-2) (as defined by COSMIN guidelines). I found no studies evaluating inter-rater reliability for PDMS-2 or QNST-3R.

Only two instruments reported intra-rater reliability (MABC-2 and TGMD-3). I found only low evidence for intra-rater reliability for the MABC-2. Two studies reported this measurement property; however, they found different results. Of note, the study reporting insufficient evidence for intra-rater reliability (Holm et al., 2013) seemed to conflate intra-rater and test-retest reliability, using different raters at two different time points to evaluate the same children. Based on this low-quality evidence, I cannot confidently rate the intra-rater reliability of this instrument. TGMD-3, on the other hand, had high-quality evidence across six studies with uniformly positive ratings.

Seven instruments had studies reporting measurement error (BOT-2, MABC-2, PDMS-2, TGMD-3, M-FUN, SOSI-M and PBS). However, to evaluate measurement error as evidence for reliability, SEM must be compared to an established MCID, as calculated by a gold-standard anchor instrument (Prinsen et al., 2018). A group of researchers examined the measurement error of BOT-2, MABC-2, and PDMS-2 (; Wuang & Su, 2009; Wuang, Su & Huang, 2012; Wuang, Su & Su, 2012). The authors calculated MCID using the Physical Task Performance Scale (PTPS) of the Chinese School Function Assessment (Hwang et al., 2004). This instrument has not been sufficiently validated to be used as an anchor; therefore, the results of the studies evaluating measurement error for these instruments are questionable. For the remaining

instruments (TGMD-3, M-FUN, PBS), no authors reported a value for MCID; therefore, I could not evaluate reliability based on measurement error for these instruments.

**Table 2.13**

*Reliability of Sensorimotor Assessments*

| Assessment | Internal Consistency | | | Test-Retest Reliability | | | Inter-Rater Reliability | | | Intra-Rater Reliability | | | Measurement Error | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | # | QoE | Rating | # | QoE | Rating | # | QoE | Rating | # | QoE | Rating | # | QoE | Rating |
| BOT-2 | 3 | H | + | 3 | H | + | 1 | VL | + | | | | 3 | H | +/-[7] |
| MABC-2 | 7 | H | +/- | 8 | H | + | 4 | H | +[5] | 2 | L | +/- | 3 | H | +[8] |
| MABC-2 Checklist | 1 | VL | +[1] | | | | 1 | VL | + | | | | | | |
| PDMS-2 | 2 | H | + | 1 | M | + | | | | | | | 2 | H | +[8] |
| TGMD-3 | 8 | H | +[2] | 5 | H | + | 6 | H | +[6] | 6 | H | + | 1 | M | ?[8] |
| M-FUN | 1 | H | +[1] | 1 | VL | + | 1 | VL | + | | | | 1 | L | ?[8] |
| TSI | | | | 1 | VL | + | 1 | VL | + | | | | | | |
| COMPS-2 | 1 | H | +[1] | 1 | VL | + | 1 | L | + | | | | | | |
| SICO | | | | 1 | VL | + | 1 | L | + | | | | | | |
| SOSI | 1 | H | +[1,3] | 1 | VL | - | 1 | L | + | | | | 1 | L | ?[8] |
| QNST-3R | 1 | H | +[1] | 1 | VL | +[4] | | | | | | | | | |
| PBS | 1 | H | + | 1 | L | + | 1 | L | + | | | | 1 | L | ?[8] |

NOTE. # = number of studies, QoE = quality of evidence across studies, H = high, M = moderate, L = low, VL = very low, + = acceptable, +/- = inconsistent, - = unacceptable

[1]High internal consistency coefficients, but no evidence of unidimensionality of constructs; internal consistency should be interpreted with caution

[2]Two studies (Simons et al., 2016; Valentini et al., 2017) showed marginally insufficient internal reliability for Ball Skills subtest of TGMD-3

[3]Adequate internal consistency reliability for ages 5-12 and 14; inadequate for age 13 – rated (+) for our population of interest

[4]Adequate test-retest reliability for ages 5-11 and 60+; inadequate for ages 12-59 – rated (+) for our population of interest

[5]One study of inadequate quality due to serious methodological concerns (Holm et al., 2013) showed insufficient inter-rater reliability

[6]One study of inadequate quality due to serious statistical concerns (Rintala et al., 2017) showed marginally insufficient inter-rater reliability

[7]MCID based on 1 point increase on Physical Task Performance Scale

[8]Studies provided SEM, but no MCID for comparison

### Validity of General Motor Instruments

Evidence for validity varies across the 12 general motor instruments. The BOT-2, MABC-2 and TGMD-3 demonstrated the most established evidence base examining validity. I can be most confident that the scores of these three instruments truly represent aspects of children's general motor skills. At the time of this review, SICO only established evidence for

content validity (this instrument is relatively new; evidence for validity is likely forthcoming). The other instruments fall somewhere in between, with adequate or strong evidence for several forms of validity.

***Content Validity.*** As noted previously, I did not use the COSMIN framework to establish content validity. Table 2.14 contains a summary of methods used by authors to establish content validity of the instruments. Of note, the quantity of methods used does not necessarily correlate to the strength of the content validity. For example, TMGD-3 has five sources of content validity, but the authors do not provide evidence of a thorough review of literature, nor do they present a theoretical model as the basis for developing their items. Nonetheless, these instruments all demonstrate numerous sources of evidence for content validity. Nearly all instruments were based on a review of literature; many also incorporated expert reviews and clinician feedback.

**Table 2.14**

*Content Validity of General Motor Instruments*

| Assessment | Content Validity |
|---|---|
| BOT-2 | B, C, E |
| MABC-2 | B, C, D |
| MABC-2 Checklist | B, C, D |
| PDMS-2 | B, C, G |
| TGMD-3 | C, D, F, H, J |
| M-FUN | A, B, D, E |
| TSI | A, B, D, E, F |
| COMPS-2 | A, B, F |
| SICO | A, B |
| SOSI-M | A, B, E, H, J |
| QNST-3R | B, C, F, J |
| PBS | A, B |

Note. A = Items derived from theoretical model; B = Literature review; C = Items revised based on feedback from users of previous versions; D = Expert panel review; E = Pilot study to revise/refine items/instructions; F = Items revised based on conventional item analyses (e.g., item discrimination, item difficulty, floor/ceiling effects, item reliability); G = Items/test structure revised based on item response theory analyses or Rasch analyses; H = Items/test structure revised based on factor analyses; I = Items subjected to readability analyses; J = Items subjected to differential item functioning analysis

***Construct Validity.*** I identified three types of construct validity of the general motor instruments: hypothesis testing validity, structural validity, and validity based on measurement invariance. Studies examining hypothesis testing validity examined convergent, known-groups, predictive, and developmental validity of the general motor instruments. Table 2.15 contains summary statistics for these forms of hypothesis testing validity (one aspect of construct validity). Nine instruments had at least one adequate- or high-quality study evaluating convergent validity with instruments that measure a similar or related construct (BOT-2, MABC-2, MABC-2 Checklist, PDMS-2, TGMD-3, M-FUN, COMPS-2, SOSI-M, and QNST-3R). Of note, for studies of convergent validity, COSMIN does not provide an exact threshold for sufficient or insufficient evidence – rather, the reviewer must evaluate if the magnitude of the relationship between the instruments makes sense based on existing literature about (1) the instruments and (2) the constructs measured by these instruments. Most studies of convergent validity demonstrated expected results (i.e., the magnitude of the relationships between instruments were as large/small as would be expected). However, a number of studies did not demonstrate expected relationships between instruments (e.g., BOT-2 vs. MABC-2; Lane & Brown, 2015). These results may suggest that the instruments measure slightly different aspects of sensorimotor ability. Alternatively, these results could call into question the construct validity of the instruments; it is possible that, for example, factors such as unclear directions or fatigue impacted children's abilities such that they performed differently on tests measuring similar constructs.

I identified studies evaluating known-groups validity for the same nine instruments (BOT-2, MABC-2, MABC-2 Checklist, PDMS-2, TGMD-3, M-FUN, COMPS-2, SOSI-M, and QNST-3R). These studies examined a variety of diagnostic groups (see footnotes of Table 2.15

for more information). Across the board, clinical groups (i.e., children with disabilities) had lower scores (or scores representing less motor ability) than typically developing children. For eight of ten instruments, all group comparisons met theoretical expectations. However, all three studies examining MABC-2 and the MABC-2 Checklist had serious methodological concerns that limited interpretation. Therefore, known groups analyses do not provide evidence of construct validity for these instruments. Furthermore, the quality of evidence scores for studies examining PDMS-2, COMPS-2 and SOSI-M were low/very low – therefore, the results should be interpreted with caution.

I identified only four instruments with studies examining predictive validity (i.e., examining scores/outcomes across time points; BOT-2, MABC-2, PDMS-2, TGMD-3). The first three instruments met hypotheses for predicting various outcomes (see Table 2.15 footnotes). For TGMD-3, a study by Wagner et al. (2017) yielded inconsistent results; TGMD-3 scores predicted ball-throwing distance but did not predict sprinting time 12 months after initial assessment. I could not locate the comparison instrument in this study (German Youth Games) for evaluation of existing psychometric properties. For the remaining instruments, I did not find studies examining predictive validity.

Six instruments had moderate or high-quality evidence for hypothesis testing validity established by examining age trends (BOT-2, PDMS-2, TGMD-3, M-FUN, SOSI-M and PBS). QNST-3R had only a single study (Mutti et al., 2017) of doubtful methodological quality. The authors did not conduct appropriate statistical tests; however, their data demonstrated convincing trends suggesting lower scores (i.e., decreases in soft signs of neurological disorder) with age, consistent with the authors' expectations.

Table 2.16 provides summary results for structural validity. Five of the instruments had

**Table 2.15**

*Construct Validity of General Motor Assessments (Hypothesis Testing)*

| Assessment | Hypothesis Testing: Convergent Validity | | | Hypothesis Testing: Known-Groups Validity | | | Hypothesis Testing: Predictive Validity | | | Hypothesis Testing: Validity based on Development | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | # | QoE | Rating | # | QoE | Rating | # | QoE | Rating | # | QoE | Rating |
| BOT-2 | 2 | M | +/-[1] | 1 | H | +[10] | 1 | M | +[19] | 1 | H | + |
| MABC-2 | 4 | H | +/-[2] | 2 | VL | ?[11] | 2 | H | +[20] | | | |
| MABC-2 Checklist | 2 | H | +/-[3] | 1 | VL | ?[12] | | | | | | |
| PDMS-2 | 5 | H | +[4] | 1 | VL | +[13] | 1 | M | +[21] | 1 | H | + |
| TGMD-3 | 1 | M | -[5] | 1 | M | +[14] | 1 | M | +/-[22] | 3 | H | +[23] |
| M-FUN | 4 | H | +[6] | 1 | H | +[15] | | | | 1 | H | + |
| COMPS-2 | 1 | M | +[7] | 1 | L | +[16] | | | | | | |
| SOSI-M | 1 | H | +/-[8] | 1 | L | +[17] | | | | 1 | H | + |
| QNST-3R | 1 | H | +[9] | 1 | H | +[18] | | | | 1 | VL | + |
| PBS | | | | | | | | | | 1 | M | + |

NOTE. # = number of studies, QoE = quality of evidence across studies, H = high, M = moderate, L = low, VL = very low, + = acceptable, +/- = inconsistent, - = unacceptable

[1]Comparison instruments = PDMS-2 (+), Test of Visual Motor Skills – Revised (+), MABC-2 (Age Band 3) (+), MABC-2 (Age Band 2) (-)

[2]Comparison instruments = BOT-2 vs. Age Band 3 (+), PDMS-2 (+), Developmental Coordination Disorder Questionnaire – Brief (DCDQ-BR) (-), TGMD-3 (-), BOT-2 vs. Age Band 2 (-)

[3]Comparison instruments = MABC-2 (+), Developmental Coordination Disorder Questionnaire (DCDQ) (+), DCDQ-BR (-)

[4]Comparison instruments = Mullen Scales of Early Learning – Adapted (MSEL:A), PDMS-2, MABC, MABC-2, M-FUN

[5]Comparison instruments = MABC-2 (-)

[6]Comparison instruments = Miller Assessment of Preschoolers (MAP), Developmental Test of Visual Perception – 2nd Edition (DVTP-2), Beery Developmental Test of Visual Motor Integration – 5th Edition (VMI-5), PDMS-2

[7]Comparison instruments = Bruininks-Osteresky Test of Motor Performance (BOTMP) and DCDQ

[8]Comparison instruments = Sensory Processing Measure (SPM), Sensory Profile – 2nd Edition (SP-2), SIPT PRN

[9]Comparison instruments = Cognitive Assessment System, Bender-Gestalt – 2nd Edition, VMI-6, BOT-2

[10]Comparison groups = Developmental coordination disorder (DCD), intellectual disability (ID), mild autism spectrum disorder (ASD)

[11]Comparison groups = Motor impairment, at-risk for motor impairment; results of both studies not interpreted due to serious methodological concerns/publishing errors (failure to describe criteria for group membership, Valentini et al., 2014; failure to include a comparison group, Henderson et al., 2007)

[12]Comparison groups = DCD, ASD; unable to interpret due to poor methodological quality of single study

[13]Comparison groups = Physical disability (unspecified)

[14]Comparison groups = ASD, ID

[15]Comparison groups = Visual motor delay, fine motor delay, gross motor delay

[16]Comparison groups = DCD

[17]Comparison groups = ASD, DCD, LD, ADHD, SPD

[18]Comparison groups = Attention deficit hyperactivity disorder (ADHD), LD, ASD

[19]Prediction variable = Physical Task Performance Scale (PTPS) subtest of School Function Assessment (SFA)

[20]Prediction variable = MABC-2; PTPS subtest of SFA

[21]Prediction variable = PTPS subtest of SFA

[22]Prediction variable = German Youth Games ball-throwing distance (+); German Youth Games sprinting (-)

[23]One small study of doubtful methodological quality showed smaller-than-expected differences between 3rd graders and 4th graders on Ball Skills subtest

articles evaluating structural validity (BOT-2, MABC-2, PDMS-2, TGMD-3, PBS) using factor analyses or Rasch analyses. Each of these instruments had at least moderate evidence for this measurement property. However, for all three of the instruments with more than one study (BOT-2, MABC-2, and TGMD-3), findings were inconsistent across authors. For each of these instruments, the instrument authors demonstrated factor analyses that supported their original test structure; however, later studies revealed alternative factor structures or rejected the authors' structures. In general, the structural validity of general motor assessments deserves further investigation.

Table 2.16 also presents evidence for cross-cultural validity/measurement invariance. I found studies examining this type of validity for four of the instruments (MABC-2, PDMS-2, TGMD-3, QNST-3R). The comparison groups for these studies varied widely, including children from different countries (e.g., Fleurkens-Peeters et al., 2019 [MABC-2]) or sex groups (e.g., Valentini et al., 2017 [TGMD-3]). There was a large degree of variability in the methodological quality of these studies. Some studies (e.g., Fleurkens-Peeters et al., 2019 [MABC-2]; Hirata et al., 2018 [MABC-2]) only provided group comparisons which are not an acceptable test of measurement invariance according to the COSMIN framework. Other studies used more rigorous methods, such as DIF (e.g., Mutti et al., 2017 [QNST-3R]) or multiple-group CFA (e.g., Magistro et al., 2018 [TGMD-3]). QNST-3R and SOSI-M had positive summary ratings for this measurement property; however, I based these ratings on only a single study (Mutti et al., 2017 and Blanche et al., 2021, respectively). The studies for MABC-2, PDMS-2, and TGMD-3 were either inconclusive due to poor methodological quality or suggested possible measurement variance among subgroups. Measurement invariance is crucial to ensuring that instruments

measure only the construct they are designed to measure; therefore, further evidence establishing this property would support construct validity for all ten instruments.

***Criterion Validity.*** Table 2.16 also includes evidence for criterion validity. Six instruments had evidence for criterion validity (MABC-2, MABC-2 Checklist, M-FUN, TSI, COMPS-2, QNST-3R). Adequate or very good evidence suggested that M-FUN, TSI, and QNST-3R could differentiate between children with and without motor delays/impairments. For MABC-2, I could not interpret the results of the single study (Valentini et al., 2014); due to methodological concerns or publishing error, the method for assigning children to groups for sensitivity/specificity analysis was unclear and appeared to be based on MABC-2 scores themselves. The single study for MABC-2 Checklist suggested inability to categorize children with, at risk for, and without motor impairments, as defined by the MABC-2 performance test (Schoemaker et al., 2012). The study evaluating criterion validity for COMPS-2 differed based on age; the instrument was adequately sensitive to detect children with disabilities under 10 years old but did not meet COSMIN criteria for children 10 and older (Wilson et al., 2000). In general, the criterion validity of the general motor instruments is inadequate. These instruments are often used to establish baseline areas of concern/strengths for children with suspected motor difficulties rather than to diagnose children; therefore, classifying children is perhaps less important. Still, misclassifying children as typically developing may prevent children from accessing services; therefore, the criterion validity of these instruments deserves further study.

**Responsiveness of General Motor Instruments**

Overall, I found very few studies examining the responsiveness of the general motor instruments. Only three instruments (BOT-2, MABC-2, and PDMS-2) had evidence evaluating this measurement property. Table 2.17 describes these studies. All three instruments were

**Table 2.16**

*Validity of General Motor Instruments*

| Assessment | Structural Validity[1] | | | Cross-cultural Validity/ Measurement Invariance[2] | | | Criterion Validity | | |
|---|---|---|---|---|---|---|---|---|---|
| | # | QoE | Rating | # | QoE | Rating | # | QoE | Rating |
| BOT-2 | 2 | H | +/- | | | | | | |
| MABC-2 | 7 | H | +/- | 3 | VL | ?[3] | 1 | VL | ?[8] |
| MABC-2 Checklist | | | | | | | 1 | H | -[9] |
| PDMS-2 | 1 | M | + | 2 | H | +/?[4] | | | |
| TGMD-3 | 6 | H | +/- | 4 | H | +/-[5] | | | |
| M-FUN | | | | | | | 1 | H | +[10] |
| TSI | | | | | | | 1 | H | +[11] |
| COMPS-2 | | | | | | | 1 | VL | +/-[12] |
| SOSI-M | | | | 1 | H | +[6] | | | |
| QNST-3R | | | | 1 | H | +[7] | 1 | M | +[13] |
| PBS | 1 | H | + | | | | | | |

NOTE. # = number of studies, QoE = quality of evidence across studies, H = high, M = moderate, L = low, VL = very low, + = acceptable, +/- = inconsistent, - = unacceptable

[1]Factor analytical studies that did not support the authors' original assessment structure (i.e., subscales) OR required substantial changes to the factor structure for adequate fit statistics were rated (-) (Prinsen et al., 2018)

[2]Cross-cultural validity/measurement invariance studies that only conducted between-group comparisons (i.e., did not conduct DIF analysis, logistic regression analysis, or multiple group factor analysis) had inadequate methodological quality and, therefore, received a rating of (?) (Prinsen et al., 2018)

[3]Comparison groups = Japanese vs. UK and Surinamese vs. UK

[4]Comparison groups = Sex (+); ethnicity (+); Portuguese vs. US (?)

[5]Comparison groups = Sex (+/-); ethnicity (+); race (+); TD vs. behavioral disorders (+); age groups (+)

[6]Comparison groups = Sex, ethnicity, hispanic origin, US region, urban-suburban vs. rural

[7]Comparison groups = Sex, race, urban-suburban vs. rural

[8]Results of the single criterion validity study (Valentini et al., 2014) not interpreted due to serious methodological concerns or publishing error (failure to describe criteria for group membership)

[9]Criterion groups = Motor impairment, no motor impairment, as determined by MABC-2 Performance Test

[10]Criterion groups = Visual motor delay, fine motor delay, gross motor delay, no motor delay, as determined by diagnosis from MD, OT, or PT

[11]Criterion groups = Unspecified delays and/or LDs

[12]Criterion groups = DCD, Non-DCD, as determined by Bruininks-Osteresky Motor Performance Test, MABC, and/or Developmental Coordination Disorder Questionnaire; authors only conducted discriminant analysis; adequate sensitivity and specificity for children < 10 years, inadequate sensitivity for children $\geq$ 10 years

[13]Criterion groups = Disability (LD or ASD), no disability, no criteria for group membership defined

[14]Methodological quality downgraded due to use of an unvalidated anchor (single point increase on Physical Task Performance Test of the School Function Assessment)

evaluated by the same research group that examined measurement errors (Wuang and colleagues). They found insufficient evidence for responsiveness using the BOT-2 and the PDMS-2; the MABC-2 had adequate responsiveness in one study examining change scores for children with DCD (Wuang, Su & Su, 2012) but inadequate responsiveness in a study of children with ID (Wuang, Su & Huang, 2012). Once again, the authors used the Physical Task

Performance Scale as the anchor score for all studies. Given the unvalidated anchor, these results are questionable. Responsiveness of the general motor instruments requires further investigation.

**Table 2.17**

*Responsiveness of General Motor Instruments*

| Assessment | # | QoE | Rating |
|---|---|---|---|
| BOT-2 | 2 | M | - |
| MABC-2 | 2 | M | +/- |
| PDMS-2 | 1 | L | - |

### General Motor Instruments: State of Measurement

The gross and fine motor instruments presented in this review of literature show a number of strengths, as well as several important limitations. For both reliability and validity, BOT-2, MABC-2, and TGMD-3 had the most empirical support. Still, studies evaluating measurement error and responsiveness for these instruments could not be easily interpreted based on the current body of literature due to the lack of valid MCID values for these instruments.

Additionally, each of these instruments lacks sufficient evidence to establish structural validity. Across studies, the BOT-2, MABC-2, and TGMD-3 demonstrated different factor structures. This finding calls into question many other measurement properties; for example, internal consistency relies upon the assumption of a valid underlying item structure.

Finally, I found very limited evidence for responsiveness of these measures, with questionable studies conducted for only three instruments. In addition to questionable psychometric properties, most of the general motor instruments are not specifically designed to evaluate sensorimotor function. While skills such as balance are reflections of discrete groups of sensorimotor function (i.e., postural function), many of the items on these tests reflect skills that involve multiple sensory systems and pathways. TSI, COMPS-2 and SICO are rooted in Ayres'

sensory integration theory; therefore, these instruments may be best suited for evaluating sensorimotor ability. However, these tests have limited evidence for validity and reliability.

**Visual Motor Instruments**

I identified four instruments designed to measure visual motor skills: Beery-Buktenica Developmental Test of Visual Motor Integration – 6th Edition (VMI-6), Developmental Test of Visual Perception – 3rd Edition (DTVP-3), Test of Visual Motor Skills – 3rd Edition (TVMS-3), and NEPSY-II (Visuomotor Precision [VP] and Design Copying [DCP]). Table 2.18 contains a brief description of each instrument. Two additional instruments (M-FUN and PDMS-2) are primarily measures of gross and fine motor, but contain visual motor scales; therefore, I included these instruments in the previous section.

All six instruments are norm-referenced. Both VMI-6 and TVMS-3 are designed for use with children and adults (2:0 – 9:11 and 3:0 – 90:0, respectively); DTVP-3, NEPSY-II, M-FUN, and PDMS-2 are appropriate for children only. The VMI-6 is primarily a measure of visual motor integration; however, the most recent edition adds optional scales to evaluate visual perception and motor coordination. DTVP-3, on the other hand, is primarily a measure of visual perception but has a visual motor component. M-FUN and PDMS-2 assess multiple motor functions. All six instruments use design copying to evaluate visual motor integration; the copying components of these tests are very similar. Copying comprises the majority of items for VMI-6, TVMS-3, and the visual motor component of DTVP-3. For NEPSY-II, the visual motor tests include a design copying section (DCP) and a visuomotor precision task (tracing a path; VP). M-FUN and PDMS-2 ask the child to demonstrate a variety of skills that require visual motor functioning (e.g., tracing a path, completing a maze, stacking blocks, and writing). None of the visual motor instruments requires extensive training.

**Table 2.18**

*Description of Visual Motor Instruments*

| Assessment | Description | Subscales Assessed | Age Range (years) | Administration Time (minutes) | Type | Training | Scores | Languages |
|---|---|---|---|---|---|---|---|---|
| Beery Developmental Visual Motor Integration Test – 6th Edition (VMI-6; Beery & Beery, 2010) | Assessment designed to evaluate visual motor skills; supplementary assessments evaluate visual perception and visual-free motor skills | - Visual motor integration<br>- Visual perception<br>- Motor coordination | 2-100 | 10-30 | Norm-referenced, performance-based | Manual review | Raw score, standard score, percentile rank, age equivalent | English |
| Developmental Test of Visual Perception – 3rd Edition (DTVP-3; Hammill, Pearson & Voress, 2014) | Developmental assessment designed to evaluate visual perception with visual motor components | - Eye-hand coordination<br>- Copying<br>- Figure-ground<br>- Visual closure<br>- Form constancy | 4-12 years | 20-40 | Norm-referenced, performance-based | Manual review | Raw score, scaled score, percentile rank | English |
| Test of Visual Motor Skills – 3rd Edition (TVMS-3; Martin, 2010) | Assessment designed to evaluate visual motor skills by asking the child to copy increasingly difficult geometric designs | None - all items assess visual motor integration | 3-90+ | 20-30 | Norm-referenced, criterion-referenced, performance-based | Manual review | Raw score, standard score, percentile rank, age equivalent | English |
| NEPSY-II (Visuomotor Precision [VP] and Design Copying [DCP]; Korkman et al., 2007) | Two tests of a longer neuropsychological evaluation; these tests examine eye-hand coordination and visual motor speed | - Visuomotor Precision (visual motor integration and Precision)<br>- Design Copying (visual motor integration) | VP: 3-12<br>DCP: 3-16 | 20-30 | Norm-referenced, performance-based | Manual Review | Raw score, scaled score, percentile rank | English |
| PDMS-2 | | | | *See Previous* | | | | |
| M-FUN | | | | *See Previous* | | | | |

### Reliability of Visual Motor Instruments

The four visual motor instruments had evidence suggesting sufficient reliability. However, most studies examining the reliability of these instruments were small and/or used inadequate statistical methods (e.g., authors calculated Pearson's *r* correlations when ICCs would have been more appropriate; Mokkink et al., 2018). Therefore, larger, high-quality studies would increase confidence in the reliability of these instruments. Moreover, none of the studies evaluate intra-rater reliability for visual motor instruments. Studies should examine the ability of raters to score these instruments consistently. Table 2.19 contains summary reliability information for the four visual motor tests.

Each test had high-quality evidence suggesting internal consistency. For VMI-6, I found no evidence for unidimensionality (a prerequisite for internal consistency) associated with the most recent edition. However, the content of the visual motor test of the VMI-6 has not changed since its original publication in 1967 (Beery, 1967). Therefore, previous studies suggesting unidimensionality of this instrument (e.g., Brown, Unsworth & Lyons, 2009) may be considered to strengthen the evidence for internal consistency.

Each test had at least one study examining test-retest reliability. For VMI-6, two studies found different results – the manual cited strong evidence for test-retest reliability for all three subtests (Beery & Beery, 2010) while an independently conducted study found insufficient evidence for this measurement property using the visual motor integration and visual perception subtests (Harvey et al., 2017). Moreover, the authors of the instrument acknowledge that test-retest reliability coefficients in previous studies have been inconsistent. DTVP-3 and TVMS-3 each have sufficient evidence for test-retest reliability, but each has only a single study. For the DTVP-3, the single study had adequate quality, but the overall body of evidence was

downgraded due to small sample size (Hammill, Pearson & Voress, 2014). For TVMS-3, the single study had doubtful methodological quality due to inappropriate statistical methods (Martin, 2010). Therefore, for all three instruments, additional high-quality studies would improve confidence in this measurement property.

Inter-rater reliability studies have been conducted for all four tests as well. VMI-6 has high-quality evidence from two studies suggesting sufficient inter-rater reliability. DTVP-3 has evidence for sufficient inter-rater reliability from one study of adequate quality; however, the quality of the evidence was downgraded due to the small sample size of this study ($N = 30$; Hammill, Pearson & Voress, 2014). TVMS-3 also had one small study of adequate quality suggesting sufficient inter-rater reliability (Martin, 2010).

The manuals for VMI-6 and DTVP-3 both list standard errors of measurement. However, both studies used Cronbach's alpha to calculate SEM; this is considered an inappropriate method in the COSMIN model (Prinsen et al., 2018). Moreover, neither study compares SEM to a value for MCID. As a result of these methodological concerns, I could not interpret measurement error for these instruments.

**Table 2.19**

*Reliability of Visual Motor Assessments*

| Assessment | Internal Consistency | | | Test-Retest Reliability | | | Inter-Rater Reliability | | | Measurement Error | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | # | QoE | Rating | # | QoE | Rating | # | QoE | Rating | # | QoE | Rating |
| VMI-6 | 2 | H | +[1] | 2 | H | +/- | 2 | H | + | 1 | VL | ?[3] |
| DTVP-3 | 2 | H | + | 1 | L | + | 1 | VL | + | 1 | VL | ?[3] |
| TVMS-3 | 1 | H | + | 1 | L | + | 1 | L | + | | | |
| NEPSY-II | 1 | H | +[1] | 1 | H | -/+[2] | | | | 1 | H | ?[3] |

NOTE. # = number of studies, QoE = quality of evidence across studies, H = high, M = moderate, L = low, VL = very low, + = acceptable, +/- = inconsistent, - = unacceptable

[1]High internal consistency coefficients, but no evidence of unidimensionality of constructs; internal consistency should be interpreted with caution

[2]DCP showed inadequate test-retest reliability; VP showed adequate test-retest reliability

[3]Studies provide SEM, but no MCID for comparison

### Validity of Visual Motor Instruments

The visual motor instruments have each been subjected to multiple studies of validity, including construct validity (i.e., hypothesis testing validity, structural validity and cross-cultural validity), criterion validity, and responsiveness. Furthermore, each instrument has several sources of evidence strengthening content validity. The DTVP-3 has the most evidence for validity, with high-quality studies examining each type of validity.

***Content Validity.*** Table 2.20 contains sources of evidence for content validity of the three instruments. Notably, the contents of these three instruments are quite similar. VMI-6 and TVMS-3 both mainly comprise design copying items, while DTVP-3 has design copying items and several other subtests examining visual perception. VMI-6 also has two supplemental tests: one examining motor control and the other examining visual perception. In each instrument's manual, the authors included a summary of existing literature supporting the use of these items. The VMI-6 presents the most thorough summary.

**Table 2.20**

*Content Validity of Visual Motor Instruments*

| Assessment | Content Validity |
|---|---|
| VMI-6 | B, G |
| DTVP-3 | B, F, J |
| TVMS-3 | B, C |
| NEPSY-II | B, C, D, E, F |

Note. A = Items derived from theoretical model; B = Literature review; C = Items revised based on feedback from users of previous versions; D = Expert panel review; E = Pilot study to revise/refine items/instructions; F = Items revised based on conventional item analyses (e.g., item discrimination, item difficulty, floor/ceiling effects, item reliability); G = Items/test structure revised based on item response theory analyses or Rasch analyses; H = Items/test structure revised based on factor analyses; I = Items subjected to readability analyses; J = Items subjected to differential item functioning analysis

***Construct Validity.*** I identified studies evaluating construct validity using hypothesis testing, structural validity, and cross-cultural/measurement invariance (see Table 2.21). Studies

evaluated three kinds of hypothesis testing validity for the visual motor instruments: convergent validity, known-groups validity, and validity based on development. The VMI-6 demonstrated convergent validity with four instruments: two handwriting assessments (Minnesota Handwriting Assessment and Test of Handwriting Skills-Revised) and one instrument of general motor and visual motor abilities (Miller Function and Participation Scales). Interestingly, the correlations between related subtests of VMI-6 and DTVP-3 were lower than expected (Brown, 2016). A previous study (Hammill et al., 2014) showed a strong correlation between DTVP-3 and VMI-5. The author suggested that small sample size and narrow age range may have contributed to the unexpected results. The manual for DTVP-3 also reports strong convergent validity with several small tests of reading and math fluency, spelling, and visual perception (Hammill et al., 2014). These relationships may be more related to the visual perception components of this instrument. I identified only one study examining convergent validity of TVMS-3 (Martin, 2010); the instrument authors found a strong correlation with VMI-4 scores. For NEPSY-II, the authors compared the VP and DCP subtests to several other instruments (see Table 2.21, Footnotes). As a rule, they found small-moderate correlations (.3 - .5) with these instruments. However, correlations were higher with expected subtests (i.e., Design Fluency on Delis-Kaplin Executive Function System). Therefore, these results support the construct validity of NEPSY-II. Further studies should investigate the relationship between these similar measures and various functional outcomes (e.g., handwriting).

Known-groups validity has been established for DTVP-3, TVMS-3, and NEPSY-II. The authors of the DTVP-3 explored scores in multiple diagnostic groups; across groups, they found expected patterns (e.g., lower scores in children with learning disabilities; Hammill, Pearson & Voress, 2014). Unfortunately, the authors provided insufficient demographic information to

compare these groups with the normative sample; therefore, I rated the overall quality of the evidence low. The author of the TVMS-3 also compared several diagnostic groups to the normative sample (Martin, 2010). These results also matched hypotheses. Finally, the NEPSY-II authors compared scores for a large number of clinical groups. In general, they found results in accordance with hypotheses; children with ASD, ID, and LDs scored lower on the visual motor subtests than their typically developing peers. For the VMI-6, I identified no studies comparing known groups. However, the manual reports studies conducted using scores generated from earlier normative data supporting known-groups validity. More studies should be conducted using the updated version of this instrument.

The manuals for three of the four instruments (VMI-6, DTVP-3, TVMS-3) reported strong evidence that the scores followed expected developmental trends. Moreover, these patterns held true across instruments: scores increased rapidly during early childhood, stabilized during middle childhood and adolescence, and remained relatively constant before decreasing in older adulthood (compared using the two instruments appropriate for testing older patients; VMI-6 and TMVS-3). These trends support the construct validity of these three instruments. I found no evidence supporting age- or development-related trends among NEPSY-II scores.

Table 2.22 contains sources of structural validity evidence. Two instruments have strong evidence of structural validity (DTVP-3 and TVMS-3) from single studies conducted by the instrument authors. While these studies support the expected subtest structure of the DTVP-3 (Hammill, Pearson & Voress, 2014) and the unidimensionality of the TVMS-3 (Martin, 2010), further studies should be conducted to confirm these findings. For the VMI-6, the authors (Beery & Beery, 2010) stated that they conducted Rasch analyses; however, they only reported person and item separation indices. While these are an important component of Rasch analyses, they

**Table 2.21**

*Construct Validity of Visual Motor Assessments (Hypothesis Testing)*

| Assessment | Hypothesis Testing: Convergent Validity | | | Hypothesis Testing: Known-Groups Validity | | | Hypothesis Testing: Validity based on Development | | |
|---|---|---|---|---|---|---|---|---|---|
| | # | QoE | Rating | # | QoE | Rating | # | QoE | Rating |
| VMI-6 | 3 | H | +/-[1] | | | | 1 | H | + |
| DTVP-3 | 2 | H | +/-[2] | 1 | L | +[5] | 1 | H | + |
| TVMS-3 | 1 | M | +[3] | 1 | M | +[6] | 1 | H | + |
| NEPSY-II | 1 | M | +[4] | 1 | M | +[7] | | | |

NOTE. # = number of studies, QoE = quality of evidence across studies, H = high, M = moderate, L = low, VL = very low, + = acceptable, +/- = inconsistent, - = unacceptable

[1]Comparison instruments = Minnesota Handwriting Assessment (+), Test of Handwriting Skills-Revised (+), Miller Function and Participation Scales (+), DTVP-3 (-)

[2]Comparison instruments = Beery VMI-5 (+), TVPS-3 (+), Test of Silent Word Reading Fluency-2 (+), Test of Silent Contextual Reading Fluency-2 (+), Test of Written Spelling-5 (+), Mathematics Fluency and Calculations Test (+), and Beery VMI-6 (-)

[3]Comparison instruments = Beery VMI-4

[4]Comparison instruments = Weschler Intelligence Scale for Children-4, Differential Abilities Scale-II, Weschler Nonverbal Scale of Ability, Weschler Individual Achievement Test-II, Children's Memory Scale, Delis-Kaplan Executive Function System, Bracken Basic Concept Scale-3, Devereux Scales of Mental Disorders, Adaptive Behavior Assessment System-II, Brown Attention Deficit Disorder Scales, and Children's Communication Checklist-2

[5]Comparison groups = Gifted, deaf/hard of hearing, speech delay, Asperger's disorder, LD, physical disability, autism, language delay

[6]Comparison groups = LD, ADHD, LD + ADHD

[7]Comparison groups = ADHD, Reading Disorder, Mathematics Disorder, Language Disorder, ID, ASD, Asperger's, Deaf/HOH, Emotional Disturbance

provide little evidence for structural validity. Therefore, more studies should examine the structural validity of the VMI-6 instrument. No studies examined the structural validity of the NEPSY-II visual motor scales.

**Table 2.22**

*Validity of Visual Motor Assessments*

| Assessment | Structural Validity[1] | | | Cross-cultural Validity/ Measurement Invariance[2] | | | Criterion Validity | | |
|---|---|---|---|---|---|---|---|---|---|
| | # | QoE | Rating | # | QoE | Rating | # | QoE | Rating |
| Beery VMI-6 | 1 | VL | ? | 1 | H | +[1] | | | |
| DTVP-3 | 1 | H | + | 1 | H | +[2] | 1 | H | +[3] |
| TVMS-3 | 1 | H | + | | | | | | |

NOTE. # = number of studies, QoE = quality of evidence across studies, H = high, M = moderate, L = low, VL = very low, + = acceptable, +/- = inconsistent, - = unacceptable

[1]Comparison groups = Sex (+)

[2]Comparison groups = Sex (+); race (+)

[3]Criterion groups = Visual perception impairment, no visual perception impairment, as determined by TVPS-3 and Beery VMI-5

Table 2.22 also describes studies of cross-cultural/measurement invariance. These studies support the construct validity of two visual motor instruments (VMI-6 and DTVP-3). While the studies had strong methodological quality, they only analyzed measurement invariance based on sex (VMI-6, Beery & Beery, 2010; and DTVP-3, Hamill, Pearson & Voress, 2014) and race (DTVP-3). Their findings supported invariance; however, further studies should examine the measurement invariance across additional groups (e.g., urban vs. rural, different language groups, etc.). No studies evaluated this measurement property for TVMS-3 or NEPSY-II.

***Criterion Validity.*** DTVP-3 had a single study evaluating criterion validity (Hamill, Pearson & Voress, 2014; see Table 2.22). Based on a ROC curve analysis, the instrument adequately specified children with and without visual perception impairments. No other studies examined this measurement property for VMI-6, TVMS-3 or NEPSY-II. Instruments must be able to identify children who need intervention; therefore, further studies should investigate the criterion validity of visual motor tests.

**Responsiveness of Visual Motor Instruments**

Only one study examined the responsiveness of visual motor instruments. Pfeiffer et al. (2015) calculated change scores on the VMI-6 for kindergarten through second-grade students who received a handwriting intervention. Based on the COSMIN standards, the study had doubtful methodological quality. Furthermore, the authors did not establish an MCID value to evaluate the responsiveness. They found that, while children in each grade level did improve on assessments of handwriting (Minnesota Handwriting Assessment, Test of Handwriting Skills-Revised) with the intervention, their scores on the VMI-6 actually decreased, although insignificantly. This calls into question the usefulness of the VMI-6 as an outcome measure. I did not identify any studies examining responsiveness for NEPSY-II, DTVP-3 or TVMS-3.

### Visual Motor Instruments: State of Measurement

Across the board, few studies provided evidence for reliability, validity and responsiveness of the four visual motor instruments. Each study had evidence for internal consistency, test-retest reliability, and inter-rater reliability; however, most studies were drawn from the instrument manuals themselves. Similarly, the evidence for validity of the visual motor instruments is somewhat weak. M-FUN and PDMS-2 have stronger evidence for validity, but these instruments are not specific to the evaluation of visual motor abilities and, therefore, may be used less frequently for this purpose. While the visual motor tests generally demonstrated strong content validity, construct validity and criterion validity lacked sufficient evidence. The instruments each had several validity studies, but most were, once again, conducted by the instrument authors themselves. The bodies of evidence supporting validity and reliability of these instruments would benefit from independent analyses conducted by authors who are not invested in the success of the instrument (Greenslade & Coggins, 2016). Furthermore, studies should evaluate the responsiveness of these instruments – this measurement property has not been sufficiently studied for any instrument.

Although the VMI-6 suffers from a similar lack of evidence as the other instruments, it should be noted that previous versions have established a strong basis for validity and reliability. Unlike other revised tests, the VMI-6 items have not changed since the original creation of the instrument; therefore, these studies (listed in the VMI-6 manual, Beery & Beery, 2010) should be considered supportive of this instrument. However, because the normative sample (and, therefore, the standardized scores) for the instrument changed, new research is critical to establishing confidence in this instrument.

The visual motor instruments are available only in English and were normed with English-speaking populations. While the design copying items on each test purport to be "culture free," I found no evidence to support this assertion. Despite this limitation, the authors present these measures as useful for international populations. Beery and Beery (2010) stated that their US-based norms have been long considered international norms, although they provide no evidence for this claim. Measurement studies should examine the invariance of these instruments across non-English speaking regions of the world.

Visual motor skills represent a critical element of sensorimotor function. Visual motor ability affords downstream occupational engagement, such as writing or driving a car. The visual motor instruments described here require less training and (with the exception of NEPSY-II), less costly materials compared to the SIPT. However, they lack sufficient evidence for validity and reliability. Therefore, these instruments provide little advantage over SIPT visual motor tests, and clinicians still lack a strong instrument for evaluating visual motor skills.

**Praxis Instruments**

I identified three alternatives to SIPT that therapists may use to evaluate praxis: Test of Ideational Praxis (TIP), Preschool Imitation and Praxis Scale (PIPS), and NEPSY-II (Fingertip Tapping [FT], Imitating Hand Positions [IH], and Manual Motor Sequences [MM]). Table 2.23 contains a brief description of each test. TIP focuses on ideational praxis: the cognitive aspect of praxis that refers to the ability to generate ideas for novel motor sequences (Lane, Ivey & May-Benson, 2014). PIPS and NEPSY-II focus on imitation – the ability to observe and imitate others' motor actions (Vanvuchelen, Roeyers & De Weerdt, 2011a).

**Table 2.23**

*Description of Praxis Instruments*

| Assessment | Description | Subscales Assessed | Age Range (years) | Administration Time (minutes) | Type | Training | Scores | Languages |
|---|---|---|---|---|---|---|---|---|
| Test of Ideational Praxis (TIP; most recent version: Lane et al., 2014)* | Single item scale that evaluates children's ability to generate novel ideas: "Show me everything you can think to do with this piece of string". Scored based on # of unique ideas. | None | 3-5[1] | 5-10 | Norm-referenced, performance-based | Required training not described | Raw scores | English (could be easily administered in other languages due to single, simple instructions) |
| Preschool Imitation and Praxis Scale (Vanvuchelen, Royers & de Weerdt, 2010a)* | Assessment designed to evaluate bodily and procedural imitation skills in young children | - Single gestural and facial imitation<br>- Sequential gestural and facial imitation<br>- Goal directed procedural imitation<br>- Non-goal directed procedural imitation | 1-5 | 10-20 | Criterion-referenced, performance-based | Required training not described | Raw scores | Dutch, English |
| NEPSY-II (Fingertip Tapping [FT], Imitating Hand Positions [IH], Manual Motor Sequences [MM]; Korkman et al., 2007) | Three tests of a larger neuropsychological evaluation; designed to evaluate imitation skills | - Fingertip tapping (finger dexterity, motor speed, rapid motor programming)<br>- Imitating hand positions (imitational praxis)<br>- Manual motor sequences (bilateral integration, imitational praxis) | FT: 5-16<br>IH: 3-12<br>MM: 3-12 | 20-30 | Norm-referenced, performance-based | Manual review | Raw scores, scaled scores, percentile ranks | English |

* = only available upon request to author

[1]Single study conducted evaluating revised TIP examines 3-5 year olds; however, previous versions have been used for children 5-8

NEPSY-II requires no training beyond reviewing assessment procedures given in the manuals/articles. The training requirements for TIP and PIPS have not been published. TIP and PIPS are cost-effective assessments, requiring little or no specialized equipment. NEPSY-II, on the other hand, is quite costly and requires a large test kit of materials. Neither TIP nor PIPS is commercially available; the authors must be contacted for full assessment, scoring protocols, and norms. None of the instruments has international norms, with TIP and NEPSY-II normed only on US children, and PIPS standardized only for children in Belgium.

As noted, many of the gross, fine, and visual motor instruments described earlier may be useful for evaluating children's abilities to plan and execute novel movements. However, this section of the review contains instruments specific to the evaluation of praxis.

**Reliability of Praxis Instruments**

In general, the evidence for reliability of the praxis instruments supports reliability (see Table 2.24); however, few studies have been conducted examining this measurement property for any of the three tests. For internal consistency, PIPS and NEPSY-II each had high-quality evidence. However, the structural validity of the praxis-related tests on NEPSY-II has not been established; therefore, the internal consistency of this instrument is questionable. Because TIP is a single-item assessment, internal consistency cannot be calculated for this instrument.

**Table 2.24**

*Reliability of Praxis Instruments*

| Assess-ment | Internal Consistency | | | Test-retest Reliability | | | Inter-rater Reliability | | | Intra-rater Reliability | | | Measurement Error | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | # | QoE | Rating | # | QoE | Rating | # | QoE | Rating | # | QoE | Rating | # | QoE | Rating |
| TIP | | | | 1 | VL | + | 1 | L | + | | | | 1 | H | ? |
| PIPS | 1 | H | + | 1 | L | + | 2 | M | + | 1 | L | + | 1 | H | ? |
| NEPSY-II | 1 | H | +[1] | 1 | M | + | | | | | | | 1 | H | ? |

NOTE. # = number of studies, QoE = quality of evidence across studies, H = high, M = moderate, L = low, VL = very low, + = acceptable, +/- = inconsistent, - = unacceptable

[1]High internal consistency coefficients, but no evidence of unidimensionality of constructs

Test-retest reliability has been established for TIP, PIPS, and the FT subscale of NEPSY-II; however, the quality of evidence varies across these instruments. Each study had adequate methodological quality, but sample sizes were too small to establish strong confidence in these results. The temporal stability of the praxis tests warrants further investigation.

Inter-rater reliability has been established for TIP and PIPS. As with test-retest reliability, the quality of evidence is negatively impacted by small sample sizes. This measurement property, too, deserves further study.

Reliability based on measurement error cannot be established for these three instruments. Although each instrument has one study publishing SEMs, no studies have generated MCID values for these instruments; therefore, the SEMs cannot be interpreted.

**Validity of Praxis Instruments**

Across the three praxis instruments, evidence for validity is somewhat sparse. While each test demonstrated good evidence for content validity, other forms of validity lack sufficient studies. For NEPSY-II, studies have only been conducted using the FT subtest. While the authors report that the IH and MM subtests have not changed since the original publication of NEPSY, this manual contains new normative data for these scales. Therefore, the validity of IH and MM should be re-examined with a new sample of children. PIPS has moderate evidence for structural validity and validity based on development. The latest version of TIP, a single-item instrument, has no evidence of validity. While the item is the same as an item from the previous version (May-Benson, 2005), new studies must reflect the single-item structure. No instruments have evidence for criterion validity.

***Content Validity.*** Table 2.25 summarizes evidence for content validity of the praxis tests. Of note, the single study evaluating the current TIP (Lane, Ivey & May-Benson, 2014) does not

describe content validity; however, I located a dissertation describing development of the original instrument (May-Benson, 2005). Each test has a firm basis in theory and literature, and at least one additional source of content validity. NEPSY-II has the most established basis for content validity. Items on NEPSY-II were revised based on theoretical and statistical information, as well as feedback from previous users.

**Table 2.25**

*Content Validity of Praxis Instruments*

| Assessment | Content Validity |
|---|---|
| TIP | A, B, E |
| PIPS | A, B, H |
| NEPSY-II | B, C, D, E, F |

Note. A = Items derived from theoretical model; B = Literature review; C = Items revised based on feedback from users of previous versions; D = Expert panel review; E = Pilot study to revise/refine items/instructions; F = Items revised based on conventional item analyses (e.g., item discrimination, item difficulty, floor/ceiling effects, item reliability); G = Items/test structure revised based on item response theory analyses or Rasch analyses; H = Items/test structure revised based on factor analyses; I = Items subjected to readability analyses; J = Items subjected to differential item functioning analysis

***Construct Validity.*** Hypothesis testing studies served as the main sources of construct validity evidence for the praxis tests. Table 2.26 contains a description of hypothesis-testing studies, including convergent validity, known-groups validity, and validity based on development/age. The NEPSY-II authors correlated results of the FT tapping test with several tests examining general cognitive ability, intelligence, academic achievement, memory, language, behavior, and comorbidities such as ADHD and mental disorders. The authors found expected correlations among subscales of these instruments and the FT subtest of NEPSY-II; as a rule, the FT had low correlations with these instruments, but modest correlations with tests requiring attention to external stimuli (i.e., recognition subtest of WNV). No studies examined convergent validity for PIPS, TIP, or other praxis-related subtests of NEPSY-II.

**Table 2.26**

*Hypothesis Testing Validity for Praxis Instruments*

| Assessment | Hypothesis Testing: Convergent Validity | | | Hypothesis Testing: Known-Groups Validity | | | Hypothesis Testing: Validity based on Development | | |
|---|---|---|---|---|---|---|---|---|---|
| | # | QoE | Rating | # | QoE | Rating | # | QoE | Rating |
| NEPSY-II | 1[1] | H | +[2] | 1[1] | H | +[3] | | | |
| PIPS | | | | | | | 1 | M | + |

NOTE. # = number of studies, QoE = quality of evidence across studies, H = high, M = moderate, L = low, VL = very low, + = acceptable, +/- = inconsistent, - = unacceptable

[1]Only FT subtest

[2]Comparison instruments = Weschler Intelligence Scale for Children-4, Differential Abilities Scale-II, Weschler Nonverbal Scale of Ability, Weschler Individual Achievement Test-II, Children's Memory Scale, Delis-Kaplan Executive Function System, Bracken Basic Concept Scale-3, Devereux Scales of Mental Disorders, Adaptive Behavior Assessment System-II, Brown Attention Deficit Disorder Scales, and Children's Communication Checklist-2

[3]Comparison groups = ADHD, Reading Disorder, Mathematics Disorder, Language Disorder, ID, ASD, Asperger's, Deaf/HOH, Emotional Disturbance

The NEPSY-II authors also completed known-groups analyses, comparing typically developing children's scores to those of children from a number of clinical groups (see Table 2.26; Footnotes). Across the board, children from clinical groups scored lower than typically developing children, with smaller differences in children with ADHD and learning disabilities. These findings align with theoretical expectations and, therefore, support the validity of NEPSY-II. However, no studies evaluated the construct validity of the IH or MM subtests of NEPSY, nor PIPS or TIP.

One study examined the progression of PIPS scores with increasing chronological and developmental age, as established by BSID-II and PDMS-2, for children with ASD. The researchers found strong evidence supporting increased scores for older and more developed children. No studies examined validity based on development/age for TIP or NEPSY-II.

Only PIPS had evidence for structural validity. One study of adequate methodological quality (Vanvuchelen, Royers & de Weerdt, 2011a) examined the factor structure of PIPS using EFA. This study suggested a four-factor model that aligned with the authors' conceptualization of imitational praxis; therefore, this supports the validity of the instrument. NEPSY-II had no

evidence of structural validity. TIP, as a single-item scale, is not appropriate for structural validity analysis.

***Criterion Validity.*** No studies have evaluated the criterion validity of the praxis tests.

**Responsiveness of Praxis Instruments**

No studies have evaluated the responsiveness of praxis instruments.

**Praxis Instruments: State of Measurement**

Praxis – the ability to plan and execute novel motor actions – provides a foundation for children's learning, exploring, and play (Ayres, 2005). Therefore, occupational therapists should be invested in evaluating and treating disorders of praxis. However, the instruments that directly assess praxis have limited benefits over the SIPT praxis tests. While these three tests are newer, and the TIP and PIPS are less expensive than SIPT, none of these tests has strong evidence for validity or reliability. Furthermore, none of these tests has evidence for responsiveness; therefore, they should not be used as valid outcome measures. PIPS and TIP each have very narrow age ranges, limiting the appropriateness of these instruments for older children. Additionally, TIP and PIPS are only available through communication with the authors – they are not available for purchase. Finally, these tests do not tap into all the dimensions of praxis ability. While SIPT includes praxis on verbal command and constructional praxis, NEPSY-II and PIPS focus only on visual imitation and TIP exclusively evaluates ideational praxis (a praxis skill that is *not* addressed on SIPT). No test, including SIPT, addresses all components of praxis.

Given the current limitations in evaluations to assess praxis skills, therapists are likely to fall back on general motor assessments or clinical observations. While these approaches do present opportunities to evaluate praxis functions, these unstandardized approaches may lead to under- or over-diagnosis of praxis problems. Thorough, praxis-oriented evaluations would allow

therapists to better evaluate these functions. SIPT are currently the most valid and reliable approach for evaluating praxis; however, for reasons described previously, SIPT are not appropriate for all clinical scenarios. Therefore, an alternative to SIPT must be created to allow therapists to evaluate praxis.

## Sensory Perception Instruments

Sensory perception refers to the ability to notice and interpret the spatial, temporal and other qualities of sensation (Bundy & Lane, 2020). Sensory perception is the foundation for higher-level sensory integrative functions, including postural, ocular, visual motor, body scheme development, and praxis. For some aspects of sensation, evaluation of perception falls outside the scope of occupational therapists and requires specialty knowledge and equipment (e.g., visual and auditory acuity). I excluded instruments for these specialty areas from this review, focusing instead on instruments appropriate for clinicians evaluating sensory integrative functions.

I identified six standardized instruments for assessing sensory perception (Table 2.27): Clinical Observations of Proprioception (COP), Motor Free Visual Perception Test – 4th Edition (MVPT-4), Test of Visual Perception Skills (TVPS-4), NEPSY-II (Arrows [AW], Geometric Puzzles [GP], Picture Puzzles [PP], and Route Finding [RF]), Developmental Test of Visual Perception – 3rd Edition (DTVP-3) and VMI-6. The majority of these instruments examine visual perception (MVPT-4, TVPS-4, NEPSY-II, DTVP-3, and VMI-6). Two of the visual motor instruments have been described in previous sections (DTVP-3, and VMI-6). Both of these tests include assessment of visual perception, although it is not their focus. One instrument examines proprioceptive perception (COP). Although COP also includes questions that pertain to general motor behavior, the focus is proprioceptive perception; therefore, I chose to include this instrument with the sensory perception items.

**Table 2.27**

*Description of Sensory Perception Instruments*

| Assessment | Description | Subscales Assessed | Age (years) | Administration Time (minutes) | Type | Training | Scores | Languages |
|---|---|---|---|---|---|---|---|---|
| Clinical Observations of Proprioception (COP; Blanche et al., 2012)* | Observational instrument in which the therapist evaluates children's behavior for indications of poor proprioceptive perception (e.g., pushing, crashing, decreased postural control) | None | 2 - 8 | 15 | Criterion-referenced, observational | Article Review | Raw score | English |
| Motor Free Visual Perception Test – 4th Edition (MVPT-4; Colarusso & Hammill, 2015) | Assessment designed to evaluate visual perception with minimal/no motor requirements | None | 4 - 80+ | 20 - 25 | Norm-referenced, performance-based | Manual Review | Raw score, standard score, percentile rank | English |
| Test of Visual Perceptual Skills – 4th Edition (TVPS-4; Martin, 2017) | Assessment designed to evaluate visual perception with minimal/no motor requirements | - Visual discrimination<br>- Visual memory<br>- Spatial Relationships<br>- Form constancy<br>- Sequential memory<br>- Visual figure ground<br>- Visual closure | 5 - 21 | 30 - 60 | Norm-referenced, performance-based | Manual Review | Raw score, standard score, percentile rank, age equivalent | English |
| NEPSY-II (Arrows [AW], Geometric Puzzles [GP], Picture Puzzles [PP], and Route Finding [RF]; | Four subtests of a longer neuropsychological evaluation; these subtests evaluate visual discrimination, form-constancy, | - Arrows (visual discrimination, spatial relationships)<br>- Geometric Puzzles (form constancy, spatial relationships)<br>- Picture puzzles | AW: 3 - 16<br>GP: 3 - 16<br>PP: 7 - 16<br>RF: 5 - 12 | 30 - 40 | Norm-referenced, performance-based | Manual Review | Raw score, scaled score, percentile rank | English |

| Assessment | Description | Subscales Assessed | Age (years) | Administration Time (minutes) | Type | Training | Scores | Languages |
|---|---|---|---|---|---|---|---|---|
| Korkman et al., 2007) | figure-ground perception, and visuospatial relationships | (figure-ground perception)<br>- Route finding (spatial relationships) | | | | | | |
| DTVP-3 | | | | *See Previous* | | | | |
| VMI-6 | | | | *See Previous* | | | | |

* = only available upon request to author

In addition, therapists use a multitude of unstandardized approaches to evaluate sensory perception, particularly in the tactile sensory modality. Because these approaches are so common in clinical settings, I chose to include them in this review (see *Unstandardized Approaches to Evaluating Sensory Perception)*.

**Reliability of Sensory Perception Instruments**

The reliability of the sensory perception measures deserves further investigation (see Table 2.28). The two visual instruments (MVPT-4 and TVPS-4) have adequate evidence to establish internal consistency and test-retest reliability. Of note, the visual perception subtest of the VMI-6 has more substantial evidence for reliability (see Reliability of Visual Motor Instruments). The COP has only very low evidence for inter-rater reliability.

**Table 2.28**

*Reliability of Sensory Perception Instruments*

| Assessment | Internal Consistency | | | Test-Retest Reliability | | | Inter-Rater Reliability | | | Measurement Error | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | # | QoE | Rating | # | QoE | Rating | # | QoE | Rating | # | QoE | Rating |
| MVPT-4 | 1 | H | +[1] | 1 | M | + | | | | 1 | VL | ?[5] |
| TVPS-4 | 1 | H | +[2] | 1 | M | + | | | | 1 | VL | ?[5] |
| NEPSY-II | 1 | H | +[3] | 1 | H | +/-[4] | | | | 1 | H | ?[5] |
| COP | | | | | | | 1 | VL | + | | | |

NOTE. # = number of studies, QoE = quality of evidence across studies, H = high, M = moderate, L = low, VL = very low, + = acceptable, +/- = inconsistent, - = unacceptable

[1]High internal consistency coefficients, but no evidence of unidimensionality of constructs; internal consistency should be interpreted with caution

[2]All subtests had sufficient evidence except "Sequential Memory" (Cronbach's alpha = .68)

[3]Study only examined internal consistency for AW, GP, and PP; structural validity has also not been established for these tests

[4]AW (-), GP (-), PP (+)

[5]Studies provide SEM, but no MCID for comparison

MVPT-4 and TVPS-4 have strong evidence for internal consistency. For MVPT-4, though, structural validity has not yet been established; therefore, the internal consistency estimates for the MVPT-4 subtests should be interpreted with caution. The TVPS-4 has strong evidence supporting structural validity, improving confidence in these findings. NEPSY-II had

strong internal consistency coefficients for three subtests (AW, GP, and PP); however, no studies examined RF. I found no evidence evaluating the internal consistency of COP.

MVPT-4 and TVPS-4 also have moderate evidence for test-retest reliability. I downgraded the strength of the evidence based on sample size ($n$ = 60 and $n$ = 71, respectively). However, for both tests, the test-retest ICCs were high. NEPSY-II had mixed evidence from one strong study examining test-retest reliability; two tests showed inadequate test-retest reliability while one (PP) was adequate. No studies evaluated test-retest reliability of COP.

One study evaluated the inter-rater reliability of COP. Blanche et al. (2012) found high inter-rater reliability. Unfortunately, the raters evaluated only four common children; therefore, the methodological quality of this study was too small to provide strong evidence for this measurement property. Of note, the visual perception instruments merely require recording the child's choice on multiple-choice type items; therefore, inter-/intra-rater reliability does not apply.

The manuals for MVPT-4, TVPS-4, and NEPSY-II report SEMs for both subtests and overall scores. As with many previous instruments, these results could not be interpreted as evidence for reliability because the authors do not compare them with a MCID value. Furthermore, both manuals reported SEMs based on Cronbach's alpha. Therefore, these SEM values should be interpreted cautiously (Prinsen et al., 2018).

**Validity of Sensory Perception Instruments**

The instruments designed to evaluate sensory perception vary in their degree of evidence for validity. Generally, COP has the lowest evidence – I only identified one study evaluating several types of validity. As with previous instruments, most studies have been conducted by

instrument authors; independent replication studies would strengthen confidence in the validity of these instruments (Greenslade & Coggins, 2016).

### Content Validity

All four tests have at least two sources of evidence for content validity. Table 2.29 describes the sources of evidence for content validity for each instrument. Generally, the MVPT-4, TVPS-4, and NEPSY-II listed more quantitative analytical approaches as evidence for content validity (e.g., DIF analyses). The authors of COP used expert review panels, literature, and Ayres' sensory integration theory to establish content validity.

**Table 2.29**

*Content Validity of Sensory Perception Instruments*

| Assessment | Content Validity |
|---|---|
| MVPT-4 | B, C, F, J |
| TVPS-4 | B, C, F, H, J |
| NEPSY-II | B, C, D, E, F |
| COP | A, B, D |

Note. A = Items derived from theoretical model; B = Literature review; C = Items revised based on feedback from users of previous/pilot versions; D = Expert panel review; E = Pilot study to revise/refine items/instructions; F = Items revised based on conventional item analyses (e.g., item discrimination, item difficulty, floor/ceiling effects, item reliability); G = Items/test structure revised based on item response theory analyses or Rasch analyses; H = Items/test structure revised based on factor analyses; I = Items subjected to readability analyses; J = Items subjected to differential item functioning analysis

### Construct Validity

Evidence for construct validity includes hypothesis testing, structural, and cross-cultural/measurement invariance validity. The sensory perception instruments have evidence for three types of hypothesis-testing validity: convergent, known-groups, and developmental. Table 2.30 summarizes this evidence. MVPT-4 and TVPS-4 have established convergent validity with each other; further studies should investigate the convergent validity of these instruments with measures designed to evaluate related constructs (e.g., reading or writing). Three subtests of NEPSY-II (AW, GP, and PP) demonstrated convergent validity with a number of related tests.

No studies investigated RF. COP demonstrated convergent validity with several instruments on SIPT; however, small sample size lowers confidence in these findings.

All four instruments have evidence for known-groups validity. MVPT-4 and TVPS-4 have high and moderate evidence, respectively, suggesting that these instruments can discriminate between children with and without disabilities that may impact visual perception. The authors of NEPSY-II only examined three of four subtests (AW, GP, and PP); however, they found strong evidence for validity in a study of adequate quality. Once again, no studies investigated RF. The authors of COP found that the instrument produced significantly different item-level scores and total scores for children with known proprioceptive problems (Blanche et al., 2012). However, the methodological quality of this study was inadequate because (1) the authors did not describe how they identified children with proprioceptive problems and (2) the authors did not provide any demographic information to compare children in the proprioceptive problem and typical groups. Further studies should be conducted to evaluate the construct validity of this instrument based on known-groups analyses.

The manuals for the MVPT-4 and TVPS- reported strong correlations between children's age and their scores. Because literature suggests that visual perception should improve with age (e.g., Bezrukikh & Terebova, 2009), these findings support the validity of data collected using these two instruments.

**Table 2.30**

*Construct Validity of Sensory Perception Instruments (Hypothesis Testing)*

| Assessment | Hypothesis Testing: Convergent Validity | | | Hypothesis Testing: Known-Groups Validity | | | Hypothesis Testing: Validity based on Development | | |
|---|---|---|---|---|---|---|---|---|---|
| | # | QoE | Rating | # | QoE | Rating | # | QoE | Rating |
| MVPT-4 | 2 | M | +[1] | 2 | H | +[5] | 1 | H | + |
| TVPS-4 | 1 | L | +[2] | 1 | M | +[6] | 1 | H | + |
| NEPSY-II | 1 | M | +[3] | 1 | M | +[7] | | | |
| COP | 1 | VL | +[4] | 1 | VL | +[8] | | | |

NOTE. # = number of studies, QoE = quality of evidence across studies, H = high, M = moderate, L = low, VL = very low, + = acceptable, +/- = inconsistent, - = unacceptable
[1]Comparison instruments = Test of Visual Perceptual Skills 3rd edition, TVPS-4
[2]Comparison instrument = MVPT-4
[3]Comparison instruments = Weschler Intelligence Scale for Children-4, Differential Abilities Scale-II, Weschler Nonverbal Scale of Ability, Weschler Individual Achievement Test-II, Children's Memory Scale, Delis-Kaplan Executive Function System, Bracken Basic Concept Scale-3, Devereux Scales of Mental Disorders, Adaptive Behavior Assessment System-II, Brown Attention Deficit Disorder Scales, and Children's Communication Checklist-2
[4]Comparison instruments = SIPT – Kinesthesia, Standing/Walking Balance; SPM – Body Awareness
[5]Comparison groups = Developmental delay, ABI, LD
[6]Comparison groups = ADHD, ASD, LD
[7]Comparison groups = ADHD, Reading Disorder, Mathematics Disorder, Language Disorder, ID, ASD, Asperger's, Deaf/HOH, Emotional Disturbance
[8]Comparison groups = Known proprioception problems

Table 2.31 contains evidence for structural validity of the sensory perception tests. Two tests have studies examining structural validity: COP and TVPS-4. TVPS-4 has high-quality evidence supporting the authors' proposed factor structure. The authors of COP conducted an EFA to examine the factor structure of the instrument. While this is considered an acceptable (although not preferred) method for establishing structural validity in COSMIN, the results should only be interpreted in comparison to a proposed or ideal factor structure (Prinsen et al., 2018). Because the authors did not hypothesize (or did not publish) an inherent factor structure in the instrument, these results cannot be interpreted to support or refute the structural validity of the instrument. No studies have examined the structural validity of MVPT-4 or NEPSY-II.

**Table 2.31**

*Validity of Sensory Perception Assessments*

| Assessment | Structural Validity | | | Cross-cultural Validity/ Measurement Invariance[2] | | | Criterion Validity | | |
|---|---|---|---|---|---|---|---|---|---|
| | # | QoE | Rating | # | QoE | Rating | # | QoE | Rating |
| COP | 1 | M | ?[1] | | | | | | |
| MVPT-4 | | | | 1 | H | +[2] | 1 | M | +[4] |
| TVPS-4 | 1 | H | + | 1 | H | +[3] | | | |

NOTE. # = number of studies, QoE = quality of evidence across studies, H = high, M = moderate, L = low, VL = very low, + = acceptable, +/- = inconsistent, - = unacceptable
[1]Authors conducted an EFA. Without a clear factor structure inherent in the test, the results of this EFA cannot be interpreted as evidence for structural validity
[2]Comparison groups = sex, race, ethnicity
[3]Comparison groups = sex, urban/rural, race
[4]Criterion groups = children with and without specific learning disability (not defined)

The manuals for MVPT-4 and TVPS-4 present high-quality evidence suggesting measurement invariance (also described in Table 2.31). For both instruments, the authors found negligible/no item bias between sex and racial groups. For MVPT-4, the authors also found no difference in item functioning for Hispanic and non-Hispanic members of the standardization sample. For TVPS-4, the authors found no measurement invariance for urban and rural participants. No studies have examined the measurement invariance of NEPSY-II or COP.

### Criterion Validity

MVPT-4 had one study examining criterion validity (see Table 2.31). Kose et al. (2019) demonstrated that the instrument could differentiate between children with and without specific learning disability (a diagnostic category that includes difficulties in language, math, listening, or other important school skills). I did not find studies examining criterion validity for TVPS-4, NEPSY-II, or COP.

## Unstandardized Approaches to Evaluating Sensory Perception

In addition to these standardized instruments, there are a number of unstandardized approaches and equipment available for sensory perception testing, especially in the tactile sensory modality. Tables 2.32 and 2.33 describe these approaches; I adapted these descriptions from Case-Smith & O'Brien (2014). Approaches in Table 2.32 provide quantitative results (e.g., monofilament testing for tactile acuity). Several of these approaches have estimated normative scores for children; however, these are still considered unstandardized because different tools (i.e., different brands of aesthesiometers) may provide slightly different results. Approaches in Table 2.33, on the other hand, do not require specialized instruments; these methods can be used in nearly any clinic. All but two tests evaluate tactile perception; the final two tests in this table

evaluate proprioception (static; when the joint is stationary, and active/kinesthetic; when the joint moves).

In general, I found few studies evaluating the validity or reliability of the assessment approaches described in Tables 2.33 and 2.34 for children ages 3-12. For static and moving two-point discrimination, Menier et al. (1996; methodological quality adequate) found adequate evidence for test-retest reliability. For vibration thresholds, Hilz et al. (1998; methodological quality adequate) found evidence to establish convergent validity between tuning forms and electric vibrometers, two instruments used to evaluate this aspect of sensory perception. They also found evidence for adequate test-retest reliability using the vibrometer.

**Table 2.32**

*Quantitative Direct Sensory Perception Assessment*

| Domain of Sensory Perception | Instrument | Stimulus (S) and Response (R) | Norms established for children? | Measurement properties established for children?[1] |
|---|---|---|---|---|
| Touch threshold/sensitivity | Monofilaments | S: Apply thinnest filament to skin until filament bends; adjust filament size using thicker filaments until the child can identify stimulus<br><br>R: Child says or signals "yes" when they feel a stimulus | Dua et al., 2019; 3 - 17 years Zingaretti et al., 2019, 6-14 years | None |
| Acuity (Static two-point discrimination) | Aesthesiometer | S: Beginning with a 5mm separation, lightly apply one or two points to the skin; adjust separation to find the smallest distance at which the child can identify correctly<br><br>R: Child indicates "one" or "two" | Dua et al., 2019; 4 - 17 years Zingaretti et al., 2019, 6 - 14 years | + Test-retest reliability; Menier et al., 1996 |
| Acuity (Moving two-point discrimination) | Aesthesiometer | S: Beginning with an 8 mm separation, lightly apply one or two points to the skin and glide across the skin; adjust separation to find the smallest distance at which the child can identify correctly<br><br>R: Child indicates "one" or "two" | Dua et al., 2019; 4 - 17 years | + Test-retest reliability; Menier et al., 1996 |
| Touch localization | Monofilaments | S: Apply filament to child's skin with vision occluded; remove<br><br>R: With vision no longer occluded, child | No | None |

| Domain of Sensory Perception | Instrument | Stimulus (S) and Response (R) | Norms established for children? | Measurement properties established for children?[1] |
|---|---|---|---|---|
| | | uses pen or finger to point to the spot touched | | |
| Vibration thresholds | Vibrometer or Tuning Fork | S: Apply vibrating head to skin, stimulus intensity is increased or decreased<br><br>R: Child indicates when vibration is felt (vibration perception threshold) or vibration is no longer felt (vibration disappearance threshold) | Hilz et al. (1998) | + Convergent validity between tuning fork and vibrometer in children ages 3 - 18 years (Hilz et al., 1998)<br>+ Test-retest reliability in children ages 3 - 7 years (Hilz et al., 1998) |

[1]+ refers to adequate measurement properties based on Prinsen et al. (2018) criteria; - refers to inadequate measurement properties

**Table 2.33**

*Additional Methods for Evaluating Sensory Perception*

| Domain of Sensory Perception | Instrument | Stimulus (S) and Response (R) | Scoring | Expected Results |
|---|---|---|---|---|
| Touch awareness | Cotton swag, fingertip, or pencil eraser | S: Light touch to a small area of skin<br><br>R: Child says or signals "yes" when they feel stimulus | # Of correct responses/# of stimuli | 100% |
| Pinprick or pain awareness | Safety pin | S: Randomly apply sharp/blunt ends of safety pin, perpendicular to skin<br><br>R: Child says "sharp" or "dull" after each stimulus | # Of correct responses/# of stimuli | 100% |
| Temperature awareness | Glass test tubes with warm and cool water | S: Apply cool or warm stimulus to skin<br><br>R: Child says "hot" or "cold" after each stimulus | # Of correct responses/# of stimuli | 100% |
| Stereognosis | Small, familiar objects (e.g., keys) | S: Place a small object in the hand<br><br>R: Child manipulates the object within the hand, names the object | # Of correct responses/# of stimuli | 100% |

| Domain of Sensory Perception | Instrument | Stimulus (S) and Response (R) | Scoring | Expected Results |
|---|---|---|---|---|
| Proprioception | None | S: Hold body segment being tested on lateral surface; move the part into different positions and hold<br><br>R: Child duplicates position with opposite extremity<br><br>OR<br><br>S: Position body segment, hold, then return to midline<br><br>R: Child returns body segment to test position | Intact, impaired, or absent | Margins of error in limb matching not established for children. Adult women able to match position within approximately 5 degrees (Kaplan et al., 1985) |
| Kinesthesia | None | S: Hold body segment being tested on lateral surface; move the part through angles of varying degrees<br><br>R: Child indicates whether part is moved up or down | Intact, impaired, or absent | 100% |

**Sensory Perception: State of Measurement**

Limited evidence supports the validity and reliability of the sensory perception instruments. While the NEPSY-II, MVPT-4 and TVPS-4 had strong psychometric properties, all but one study (Kose et al., 2019) came from instrument authors; independent studies should be conducted to strengthen the evidence base for these instruments (Greenslade & Coggins, 2016). Visual perception instruments reviewed in previous sections (DTVP-3, and VMI-6) have similar weaknesses.

Aside from visual perception, therapists have few standardized, valid, and reliable instruments useful for measuring sensory perception. There are no adequate standardized measurement approaches to evaluate proprioceptive perception. Although COP is primarily a measure of proprioceptive perception, the test examines several other constructs including

postural control, muscle tone, and motor planning. These items may be selected to represent the observable manifestations of poor proprioceptive perception. However, the focus of this test remains unclear and, therefore, COP should be used with caution. Tactile tests are largely unstandardized, and many require specialized equipment. Therefore, the results of these tests are also of questionable use to therapists. I did not identify tests examining auditory perception; however, this is usually the domain of audiologists rather than occupational therapists.

The ability to perceive and interpret sensory information is thought to be a critical foundation for sensorimotor/praxis abilities, as well as higher-level skills such as complex thinking and social interaction (Bundy & Lane, 2020). When a child has deficits in sensory perception, these higher-level functions may be impaired. While some deficits in sensory perception are very clear (e.g., blindness), others are more subtle (e.g., proprioceptive deficits). Therefore, therapists must have valid and reliable assessment tools that can evaluate sensory perception.

### Sensory Reactivity Instruments

Sensory reactivity refers to the degree of responsiveness to the intensity and duration of sensation (Bundy & Lane, 2020). SIPT do not include formal evaluations of sensory reactivity; however, more recent models of sensory integration acknowledge this factor (e.g., Bundy & Lane, 2020; Dunn et al., 2014). Children with deficits in sensory reactivity may present with over-, under-, or fluctuating responsiveness to different types of stimuli (or in different contexts, [e.g., at school or home]).

I identified four instruments useful for the evaluation of sensory reactivity: Sensory Profile – 2nd Edition (SP-2), Short Sensory Profile (SSP-2), Sensory Processing Measure (SPM)[1], and Sensory Experiences Questionnaire – 3rd Edition (SEQ 3.0). Table 2.34 describes these instruments. All three instruments rely upon proxy report, gathering perspectives from teachers, home caregivers, or both. SP-2 comprises five forms: Infant, Toddler, Child, Short, and School Companion. For this review, I excluded Infant and Toddler forms, as they were not germane to the target population (ages 3-12). In the tables that follow, I combined the school companion and child version, as most psychometrics for both versions were conducted using the same population. I separated SSP-2 because this shorter version is often used independently in both research and clinical settings, and several research studies addressed only the short form.

SPM also contains multiple forms: the Home, Main Classroom, and School Environments forms. As with SP-2, I treated these as a single instrument. When studies investigated only one form, I indicated this in a footnote.

[1] Since the original submission of this proposal, an updated version of SPM has been published (SPM-2). I could not access psychometric information for this test in time for submission of this dissertation; however, if I move forward with publication of this review, I will replace SPM with SPM-2.

**Table 2.34**

*Sensory Reactivity Instruments*

| Assessment | Description | Subscales Assessed | Age Range (years) | Administration Time (minutes) | Type | Training | Scores | Languages |
|---|---|---|---|---|---|---|---|---|
| Sensory Profile – 2nd Edition (SP-2; Dunn, 2014) | Survey designed to evaluate sensory reactivity patterns in everyday contexts | - Modalities (auditory, visual touch, movement, body position, oral sensory)<br>- Behavioral (conduct, social emotional, attentional)<br>- Responsivity (seeking, avoiding, sensitivity, registration) | 3-14 | 15-20 | Norm-referenced, caregiver or teacher report | Manual review | Raw score, standard score, percentile rank | English, Spanish, Chinese |
| Short Sensory Profile – 2nd Edition (SSP-2; Dunn, 2014) | Shortened version of SP-2 | - Modalities (auditory, visual touch, movement, body position, oral sensory)<br>- Behavioral (conduct, social emotional, attentional)<br>- Responsivity (seeking, avoiding, sensitivity, registration) | 3-14 | 5-10 | Norm-referenced, caregiver or teacher report | Manual review | Raw score, standard score, percentile rank | English, Spanish, Chinese |
| Sensory Processing Measure (SPM; Parham et al., 2007) | Survey designed to evaluate children's sensory integrative functions with a focus on sensory reactivity. SPM also contains questions related to motor/praxis. School form available. | - Social participation<br>- Vision<br>- Hearing<br>- Touch<br>- Body awareness<br>- Balance and motion<br>- Planning and ideas<br>- Total Sensory Systems | 5-12 | 15-20 | Norm-referenced, caregiver or teacher report | Manual review | Raw score, standard score, percentile rank | English, Danish, Finnish, Spanish |

| Assessment | Description | Subscales Assessed | Age Range (years) | Administration Time (minutes) | Type | Training | Scores | Languages |
|---|---|---|---|---|---|---|---|---|
| Sensory Experiences Questionnaire (SEQ 3.0; Ausderau et al., 2014)* | Survey designed to evaluate patterns of children's sensory responsivity | - Modalities (auditory, visual, tactile, gustatory/ olfactory, vestibular proprioceptive)<br>- Contexts (social, non-social)<br>- Responsivity (hyporeactive; hyperreactive; sensory interests, repetitions, and seeking behaviors; enhanced perception) | 2-12 | 15-20 | Criterion-referenced, caregiver report | Training not described | Raw score | English |

* = only available upon request to author

**Reliability of Sensory Reactivity Instruments**

Three of the four instruments (SP-2, SSP-2, and SPM) have small bodies of evidence evaluating reliability (see Table 2.35). The bodies of evidence for reliability-related measurement properties varied across studies, often due to small sample sizes or questionable statistical approaches. No evidence has examined reliability of data collected using SEQ 3.0. In general, all four instruments need additional evidence examining reliability.

I excluded intra-rater reliability in this section because all three instruments rely upon proxy report. Intra-rater reliability would require the evaluator to complete the instrument on two occasions – this could be conflated with test-retest reliability. Furthermore, I found no studies for any instrument purporting to evaluate this property.

SP-2, SSP-2, and SPM all had high-quality evidence examining internal consistency. For SP-2 and SSP-2, this evidence supported reliability. For most subtests of SPM, internal consistency evidence was sufficient; however, two studies (Brown et al., 2010b; Lai et al., 2011) suggested that the Taste and Smell subscale of the home form was not internally consistent, and one study (Lai et al., 2011) showed that the Balance and Motion subscale of the home form was not internally consistent. These internal consistency estimates may be impacted by the small number of items on these subscales. Additionally, our understandings of the role of the olfactory and gustatory sensory systems in sensory integration theory are underdeveloped (Bundy & Lane, 2020); therefore, the items on the Taste and Smell subscale may not be sufficient.

While SP-2, SSP-2 and SPM all had evidence supporting sufficient test-retest reliability, the quality of this evidence varied. SP-2 had only one study from the manual examining this measurement property, but the methodological quality was low due to very inconsistent retest intervals without sufficient justification of this range (Dunn, 2014). SSP-2 had two studies: one

from the manual, derived from the SP-2 study (Dunn, 2014) and one small but adequate study conducted by an independent research group using a Polish version (Chojnicka et al., 2019). SPM had the strongest evidence, with two adequate studies both supporting test-retest reliability.

I identified only two studies examining the inter-rater reliability of the sensory reactivity instruments (SP-2 and SPM). The single study for the SP-2 showed adequate inter-rater reliability between children's caregivers; however, the sample size was small, and the authors did not use ICCs, resulting in a low rating for the body of evidence (Dunn, 2014). Brown et al. (2010b) examined the inter-rater reliability between mothers and fathers; this study suggested insufficient reliability across caregivers. Neither SSP-2 nor SEQ 3.0 had any evidence evaluating this measurement property.

SP-2, SSP-2, and SPM reported standard error of measurement in their test manuals. As with the sensorimotor instruments, the authors did not provide MCID values to compare the SEMs. Therefore, these results cannot be interpreted. I found no SEM values for the SEQ 3.0. Of note, the manual for this instrument is only available upon request to the authors; therefore, the SEMs may be calculated but were unavailable at the time of this review.

**Table 2.35**

*Reliability of Sensory Reactivity Assessments*

| Assessment | Internal Consistency | | | Test-retest Reliability | | | Inter-rater Reliability | | | Measurement Error | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | # | QoE | Rating | # | QoE | Rating | # | QoE | Rating | # | QoE | Rating |
| SP-2 | 1 | H | + | 1 | L | + | 1 | L | + | 1 | H | ?[3] |
| SSP-2 | 2 | H | + | 2 | M | + | | | | 1 | H | ?[3] |
| SPM | 4 | H | +/-[1] | 2 | H | + | 1[2] | M | - | 1 | H | ?[3] |

NOTE. # = number of studies, QoE = quality of evidence across studies, H = high, M = moderate, L = low, VL = very low, + = acceptable, +/- = inconsistent, - = unacceptable

[1]Two studies showed inadequate internal consistency for the Taste and Smell subscale (Home form); one study showed inadequate internal consistency for Balance and Motion subscale (Home form)

[2]Single study evaluated the SPM – Home form only (Brown et al., 2010b)

[3]Studies provide SEM, but no MCID for comparison

## Validity of Sensory Reactivity Instruments

As with reliability, evidence for validity varies across the four instruments evaluating sensory reactivity. Both SSP-2 and SPM had multiple high-quality sources of evidence supporting validity, while SP-2 and SEQ 3.0 each had only two studies examining validity.

### Content Validity

Table 2.36 contains sources of evidence for content validity of the four sensory reactivity instruments. The instruments are all derived from various models of sensory integration theory, supporting their construct validity. Additionally, all authors conducted thorough reviews of literature to select representative items. The authors each took additional steps to ensure that the items adequately conveyed the constructs they intended to measure. SEQ 3.0 has the most limited evidence for content validity, while SP-2 and SSP-2 have the most.

**Table 2.36**

*Content Validity of Sensory Reactivity Instruments*

| Assessment | Content Validity |
|---|---|
| SP-2 | A, B, C, D, E, F, G, I |
| SSP-2 | A, B, C, D, E, F, G, I |
| SPM | A, B, D, F, G |
| SEQ 3.0 | A, B, C |

Note. A = Items derived from theoretical model; B = Literature review; C = Items revised based on feedback from users of previous versions; D = Expert panel review; E = Pilot study to revise/refine items/instructions; F = Items revised based on conventional item analyses (e.g., item discrimination, item difficulty, floor/ceiling effects, item reliability); G = Items/test structure revised based on item response theory analyses or Rasch analyses; H = Items/test structure revised based on factor analyses; I = Items subjected to readability analyses; J = Items subjected to differential item functioning analysis

### Construct Validity

Tables 2.37 and 2.38 review sources of construct validity evidence for the sensory reactivity tests. Studies examine two types of hypothesis-testing validity for SP-2, SSP-2, and SPM. Table 2.37 displays summary results of these studies. These three instruments each had

evidence for convergent validity. SPM had strong evidence from four studies suggesting convergent validity with previous versions of the Sensory Profile. Both SP-2 and SSP-2 had mixed evidence for convergent validity. The SP-2 manual (which also contains data for SSP-2, derived from the long form) suggests that the instruments converge with the Behavior Assessment for Children (2nd Edition; BASC-2) and the Social Skills Improvement Rating System (SSIS), but the scores were very weakly correlated with the Vineland Adaptive Behavior Scale (2nd edition; VABS-2). While the authors expected a higher correlation, the weak relationship with VABS-2 may be explained by the myriad factors that influence adaptive behavior. While sensory integration may play a role in a child's functional abilities, environmental factors, temperament, and social factors may also influence this construct. An additional study also demonstrated a convergent relationship between scores on SSP-2 and Social Communication Questionnaire (SCQ), supporting the validity of the former instrument.

**Table 2.37**

*Construct Validity of Sensory Reactivity Assessments (Hypothesis Testing)*

| Assessment | Convergent Validity | | | Known-groups Validity | | |
|---|---|---|---|---|---|---|
| | # | QoE | Rating | # | QoE | Rating |
| SP-2 | 1 | M | +/-[1] | 1 | M | +[4] |
| SSP-2 | 2 | H | +/-[2] | 2 | H | +[4] |
| SPM | 4 | H | +[3] | 3[5] | H | +[6] |

NOTE. # = number of studies, QoE = quality of evidence across studies, H = high, M = moderate, L = low, VL = very low, + = acceptable, +/- = inconsistent, - = unacceptable
[1]Comparison instruments = BASC-2 (+), SSIS (+), VABS-2 (-)
[2]Comparison instruments = BASC-2 (+), SSIS (+), VABS-2 (-), SCQ (+)
[3]Comparison instruments = Sensory Profile (1st edition), Sensory Profile School Companion (SPSC), SSP, Chinese Sensory Profile
[4]Comparison groups = DD, ASD, ADHD, ADHD and ASD, LD, gifted and talented
[5]Two studies examined SPM-Home only (Lai et al., 2011; Alkhalifah et al., 2020)
[6]Comparison groups = ASD, FASD

In addition to convergent validity, SP-2, SSP-2 and SPM each had at least moderate evidence supporting known-groups validity. For each of these instruments, children with known concerns in sensory reactivity (e.g., ASD) demonstrated significantly more hyper-/hypo-

responsiveness on all three instruments. These findings support construct validity for these three instruments. No studies examined hypothesis testing validity for SEQ 3.0.

Table 2.38 contains additional sources of construct validity evidence. Three instruments (SSP-2, SPM, and SEQ 3.0) had studies examining structural validity. Both the SSP-2 and the SEQ 3.0 had positive ratings, while SPM showed inconsistent results; however, the SPM is the only instrument with more than one study examining structural validity. The two studies (Parham et al., 2007; Lai et al., 2011) found different factor structures. Given the inconsistency, the structural validity of this instrument remains questionable. Additionally, further studies should be conducted to support the structural validity of SSP-2 and SPM. Furthermore, the structure of SP-2 should be examined.

**Table 2.38**

*Validity of Sensory Reactivity Instruments*

| Assessment | Structural Validity[1] | | | Cross-cultural Validity/ Measurement Invariance[2] | | | Criterion Validity | | |
|---|---|---|---|---|---|---|---|---|---|
| | # | QoE | Rating | # | QoE | Rating | # | QoE | Rating |
| SSP-2 | 1 | H | + | | | | | | |
| SPM | 2 | H | +/- | | | | 2 | H | +[5] |
| SEQ 3.0 | 1 | H | + | 1 | H | +[4] | | | |

NOTE. # = number of studies, QoE = quality of evidence across studies, H = high, M = moderate, L = low, VL = very low, + = acceptable, +/- = inconsistent, - = unacceptable

[1]Factor analytical studies that did not support the authors' original assessment structure (i.e., subscales) OR required substantial changes to the factor structure for adequate fit statistics were rated (-) (Prinsen et al., 2018)

[2]Cross-cultural validity/measurement invariance studies that only conducted between-group comparisons (i.e., did not conduct DIF analysis, logistic regression analysis, or multiple group factor analysis) had inadequate methodological quality and, therefore, received a rating of (?) (Prinsen et al., 2018)

[3]Comparison groups = ASD vs. ADHD

[4]Comparison groups = Sex, age groups

[5]Criterion groups = SI dysfunction, ASD

Only one study examined measurement invariance. Ausderau et al. (2014) used multiple-group CFA to establish invariance of the SEQ 3.0 based on sex and age groups. No other instruments had any evidence for this measurement property. Invariance is a critical element of

measurement (Bond, Yan & Heene, 2020); without evidence of this property, the construct validity of these instruments remains in doubt.

### Criterion Validity

Table 2.38 also reviews criterion validity evidence for the sensory reactivity instruments. Two studies investigated the criterion validity of SPM (Parham et al., 2007; Alkhalifah et al., 2020). These studies found that the SPM could distinguish between children with and without sensory processing difficulties and children with and without autism. No studies examined the criterion validity of SP-2, SSP-2, or SPM.

## Responsiveness of Sensory Reactivity Instruments

No studies have investigated the responsiveness of sensory reactivity instruments to change due to maturation or intervention.

## Sensory Reactivity Instruments: State of Measurement

Few studies have evaluated the psychometric properties of data collected using the sensory reactivity instruments. While the SP-2, SSP-2, and SPM each have several sources of evidence examining reliability and validity, the overall size of the body of evidence is small. Many studies are drawn directly from the instrument manuals. Most studies have excellent or adequate methodological quality; however, independent replication studies would improve confidence in the measurement properties of these instruments (Greenslade & Coggins, 2016). SEQ 3.0 has very little evidence in general, with no studies examining reliability. Although this is not a widely used instrument, SEQ 3.0 has been used in research studies to examine patterns of sensory reactivity for children with autism (e.g., Ausderau, 2014); therefore, the instrument deserves further investigation.

None of the sensory reactivity instruments has evidence supporting responsiveness. This measurement property is critical for establishing valid outcome measures. Sensory reactivity measures are not frequently used as outcome measures; rather, they allow therapists to establish a baseline and identify areas for intervention. However, because these instruments have been used to evaluate interventions (e.g., Hall & Case-Smith, 2007), studies should be conducted to evaluate their responsiveness to change.

The sensory reactivity instruments all rely upon caregiver report. While perspectives from caregivers are essential components of a well-rounded assessment of sensory integrative functions, inter-rater reliability studies suggest that these perspectives can vary across caregivers and contexts (e.g., Brown et al., 2010b). A clinician-rated, observational instrument evaluating sensory reactivity would add a critical element to the pool of available instruments.

### Outcomes of the Literature Review: A Need for a Novel Instrument

As this literature review demonstrates, there is a paucity of valid and reliable standardized instruments useful for measuring sensory integration functions. The current gold-standard suite of tests, SIPT, is wrought with problems including questionable validity, outdated norms, and a lack of clinical utility. As a result of these limitations, therapists may fall back on alternative instruments.

I identified 26 instruments useful for evaluating aspects of sensorimotor function. I divided these instruments into three groups: general (gross/fine) motor, visual motor, and praxis. Evidence for validity and reliability varied immensely across these instruments, with some (e.g., BOT-2) showing strong psychometric properties and others (e.g., PIPS) having very limited evidence. Praxis instruments showed the most limitations; very few instruments explicitly measured aspects of praxis, and those that did showed weak or underdeveloped measurement

properties. Praxis is fundamental for engagement in everyday occupations (Cermak & May-Benson, 2020); therefore, this lack of assessment tools presents a critical gap in existing measurement of sensory integration.

I found four instruments useful for evaluating sensory perception. Three of these instruments examined visual perception; only one instrument evaluates proprioceptive perception. I found no standardized instruments useful for evaluating tactile perception – therapists mainly use unstandardized approaches. Other aspects of sensory perception (e.g., auditory) may not be evaluated at all in most occupational therapy clinics. The ability to perceive and interpret sensation in all modalities provides a basis for sensorimotor and praxis abilities (Lane, 2020). When sensory perception is dysfunctional, children may struggle to engage with their environments and occupations. For example, a child without a strong sense of body scheme (i.e., proprioceptive perception) may struggle to dress themselves or sit still in a chair during class. Therefore, further assessment into sensory perception should be important to occupational therapists.

Finally, I identified four standardized instruments useful for evaluating sensory reactivity. While several of these instruments demonstrated adequate measurement properties (e.g., SPM), all four rely upon caregiver report. Sensory reactivity ratings appear to vary significantly across different caregivers and in different contexts (Brown et al., 2010b; Ausderau et al., 2014); therefore, these instruments may not produce a complete picture of sensory reactivity. Clinician-rated instruments that produce valid and reliable data would be a powerful addition to the available sensory reactivity assessments.

This review highlights the need for a single instrument that will produce valid and reliable data about all the constructs in sensory integration. While this dissertation (and,

therefore, this review) focuses primarily upon psychometric properties, the novel instrument must also be clinically usable, normed internationally to apply to children across the globe, and cost effective. In this dissertation, I evaluated the evidence for validity and reliability of the Evaluation in Ayres Sensory Integration (EASI): a novel instrument that has the potential to close the critical gaps in available assessment tools highlighted by this review of literature.

# CHAPTER 3: MANUSCRIPT 1: PRAXIS

In sensory integration (SI) theory, praxis is the process of conceptualizing and planning motor actions using sensory information gathered from objects, the environment, and the body's position in space (Ayres, 1989; Cermak & May-Benson, 2020). For children, praxis is a critical component of learning, behavior, and engagement with everyday occupations. When praxis does not develop appropriately (i.e., dyspraxia), children may struggle with important areas of occupation such as self-care, school tasks, and social participation (Engel-Yeger, 2015; Izadi-Najafabadi et al., 2019 Magalhaes, Cardoso & Missiuna, 2011).

While a variety of disciplines and theories address motor incoordination, SI theory is unique in considering the joint roles of sensation and movement in the development of praxis. In SI theory, praxis requires awareness of body position in space; this is driven by information from various sensations: tactile, proprioceptive, visual, vestibular, and auditory (Cermak & May-Benson, 2020). Furthermore, children must *also* be able to understand the characteristics of objects and environments and to use this understanding to generate ideas for completing motor tasks (*ideation*). Dyspraxia, therefore, may be related to an inability to integrate and usefully apply sensory information.

Sensory integration theory also provides a framework for treatment of children with dyspraxia. During SI therapy, specially trained occupational therapists collaborate with the child to identify play-based sensorimotor activities that present "just-right challenges", prompting adaptive responses (i.e., an appropriately timed, graded and performed motor action in response to sensory stimuli) (Ayres, 1979; Parham et al., 2011).

## Sensory Integration Intervention Depends on Assessment

SI therapy relies on evaluation of children's unique strengths and challenges. Without proper assessments, therapists may fail to optimally design environments and tasks to promote adaptive responses and development of more advanced praxis skills. Further, a lack of standardized outcome measures makes it difficult to evaluate the impacts of SI intervention on praxis abilities; therefore, it is difficult for therapists to adjust their treatment protocols when the child is not responding.

Currently, the Sensory Integration and Praxis Tests (SIPT; Ayres, 2005) are the gold-standard for assessing praxis skills. This suite of assessments contains 5 tests evaluating praxis: Postural Praxis (PPr), Praxis on Verbal Command (PrVC), Constructional Praxis (CPr), Sequencing Praxis (SPr), and Oral Praxis (OPr). However, SIPT have several shortcomings. First, they were developed and normed over 40 years ago, and the normative sample only included children from North America. Therefore, the generalizability of these norms is questionable. Second, SIPT require proprietary materials and a costly scoring program, limiting accessibility to many underfunded clinics. Third, reliability and validity of data collected using SIPT lack sufficient evidence, especially for test-retest reliability, convergent validity and known-groups validity (see Ayres, 2005). Sample sizes used in the SIPT analyses were small and lacked diversity. No studies evaluated responsivity, limiting the use of SIPT as an outcome measure. Finally, SIPT omit ideational praxis and thus, no widely used standardized tests measure this cognitive component of praxis.

Given the limitations of SIPT, therapists evaluating children with possible praxis concerns rely heavily on widely-available developmental motor evaluations (e.g., the Movement Assessment Battery for Children – 2nd Edition [MABC-2; Henderson et al., 2007] and the

Bruininks-Osteresky Test of Motor Performance – 2nd edition [BOT-2; Bruininks & Bruininks, 2005]). While these tests allow examiners to observe many aspects of praxis, they do not align directly with SI theory (as the SIPT does). Test results primarily reflect gross and fine motor abilities rather than praxis or motor planning; importantly, these tests do not incorporate the sensory components of motor behavior (e.g., tactile discrimination). While some items require praxis, children are only scored upon *execution* of the motor abilities. Despite these limitations, skilled clinicians may (and often do) observe other aspects of praxis (i.e., planning) during the motor items. As a result, evaluating praxis with motor instruments means that the results are only as valid as the clinician's skill and knowledge about praxis.

Several instruments other than SIPT do evaluate specific aspects of praxis. The Test of Ideational Praxis (TIP; May-Benson, 2005) evaluates children's ability to generate and execute ideas for motor actions. The Preschool Imitation and Praxis Scale (PIPS; Vanvuchelen et al., 2011) focuses on imitation praxis – the ability to watch and recreate motor actions/sequences. While each of these scales provide some measurement of praxis, neither provides a complete picture, and neither is standardized with a large group of children. NEPSY-II (Korkman, Kirk & Kemp, 2007) also evaluates imitation praxis. While this instrument is standardized, it is not grounded in sensory integration theory and, like BOT-2 and MABC-2, provides limited insight into sensory foundations of dyspraxia.

**The Evaluation in Ayres Sensory Integration**

The Evaluation in Ayres Sensory Integration (EASI; Mailloux et al., 2018) is a novel, norm-referenced, performance-based suite of instruments used to assess praxis, sensory perception and sensory reactivity in children ages 3-12 years. The EASI comprises 20 individual tests; seven of these tests represent aspects of praxis. Table 1 describes the tests and scoring

procedures. Praxis Positions (PrP), Praxis Sequences (PrS) and Ocular Praxis (PrOc) evaluate praxis by observing the child's ability to replicate novel static and sequential positions of the body, hands, face and eyes. Praxis Following Direction (PrFD) evaluates the ability to plan motor actions based on verbal commands. Visual Praxis Designs (VPrD) and Visual Praxis Construction (VPrC) evaluate the ability to plan motor actions based on visual input. For VPrD, the child copies printed designs that are increasingly complex. For VPrC, the child recreates a static scene based on a visual model (a photograph)[2]. Finally, the Praxis Ideation (PrI) test evaluates ideational praxis, by asking children to demonstrate ideas for using/interacting with (1) a chair, (2) their hands, (3) a group of small everyday objects, and (4) their body. Together, these tests reflect a thorough operationalization of praxis. These tests can be used in conjunction with the 8 EASI tests evaluating sensory perception to (1) thoroughly evaluate children for dyspraxia and (2) strengthen theoretical understandings of sensory-based dyspraxia.

[2]Originally, VPrC contained additional tabletop items designed to measure visual praxis (tearing and folding paper, completing simple puzzles). These were included in the normative data collection procedure. However, after an expert panel reviewed these items, our team decided that they do not measure the same construct as the rest of the items; therefore, they are not included in these analyses.

**Table 3.1**

EASI Praxis Tests

| Test | Description | Scoring | # of Items |
|---|---|---|---|
| Praxis: Positions (PrP) | Participant imitates static positions demonstrated by the examiner | 2: Correct position<br>1: Correct position, but slight variation[1]<br>0: Incorrect position | 24 |
| Praxis: Sequences (PrS) | Participant imitates sequences of positions demonstrated by the examiner | 2: Completes all positions correctly<br>1: 1-2 errors in position or sequence[1]<br>0: More than 2 errors or extra actions | 27 |
| Ocular Praxis (PrOc) | Participant imitates eye movements demonstrated by the examiner | 2: Fluid movement, correct position<br>1: Poor fluidity but correct position<br>0: Incorrect position | 8 |
| Praxis: Following Directions (PrFD) | Participants execute static positions based on verbal instructions | 1: Correct position[2]<br>0: Incorrect position | 18 |
| Visual Praxis: Designs (VPrD) | Participants copy two-dimensional designs; some items have a dot grid while others are free-handed | 2: Correct design<br>1: Mostly correct design with slight variation[1]<br>0: Incorrect design | 24 |
| Visual Praxis: Construction (VPrC) | Participants arrange objects/small furniture to make a "silly room" based on a visual model (photograph) | 1: Object is correctly positioned or oriented<br>0: Object is not correctly positioned or oriented | 30 |
| Praxis Ideation (PrI) | Participants demonstrate all the actions they can think to do in 60 seconds using their bodies, hands, small objects, and a chair | Tally<br>3: 12+ actions<br>2: 8 – 11 actions<br>1: 4 – 7 actions<br>0: 0 – 3 actions<br><br>Variety<br>2: Many different types of actions<br>1: Several different types of actions | 12 (Body, Hands, Objects and Chair each scored for Tally, Variety and Complexity) |

| Test | Description | Scoring | # of Items |
|---|---|---|---|
| | | 0: Almost all types of actions are identical or similar<br><br>Complexity<br>2: Many creative, imaginary and/or complex actions<br>1: Some creative or complex types of actions<br>0: Mostly simple/conventional types of actions | |

[1]Examples of slight variations/errors are provided for each item

[2]This test was originally scored on a 2, 1, 0 score where 1 represented slight variations to the intended position; however, the results of the Rasch analysis completed in this study revealed that the collapsed 1, 0 scale better fit the data

**Establishing Psychometric Properties of the EASI Praxis Tests**

The EASI Praxis tests present a much-needed contribution to the science of dyspraxia and sensory integration. However, before these tests can be used, they must be subjected to rigorous psychometric analysis to establish evidence for validity and reliability. Our research group evaluated four of the praxis tests (PrP, PrS, PrFD, and PrI; Mailloux et al., under review) in a US-based pilot study. We found strong evidence for construct validity and reliability of data analyzed using the Rasch measurement model (adequate fit statistics and reliability coefficients; evidence that total test scores discriminated between children with and without known SI dysfunction; Mailloux et al., under review). However, the tests have been revised and shortened significantly since the completion of this study. Additionally, this data only includes children from the United States – a larger, international sample would strengthen confidence in the results.

In a more recent study, researchers (including myself) examined these four tests with an Israeli sample of typically developing children (Lamash et al., 2022). Cronbach's alpha revealed moderate to excellent evidence for internal consistency of the four tests *($\alpha$ = .55 - .83)*. Furthermore, this study demonstrated a positive and significant correlation between child age and EASI scores, supporting the construct validity of these tests. Although the analyses were conducted using the revised tests based on Mailloux et al. (under review), the children completed all original items from the 2018 pilot version. The extra length of the tests may have impacted children's total scores and, therefore, study results.

In 2018, a team of trained data collectors (primarily occupational therapists who volunteered to be part of data collection) began collecting EASI datasets using the revised tests with typically developing children around the world. These data will provide normative values

against which clinicians will evaluate children with dyspraxia and other sensory integration concerns. The stakes of these normative data are high – clinicians will use the EASI (in conjunction with clinical observations and interviews) to identify children who will benefit from sensory integrative treatment (Mailloux et al., 2018). Therefore, it is critical to establish the validity and reliability of normative data collected using the EASI tests.

In this study, I conducted Rasch analyses of data collected with the seven EASI praxis tests. Rasch analysis provides evidence for construct validity (i.e., the items measure the constructs they are designed to assess (Bond, Yan & Heene, 2020). I evaluated each test seperately. Specifically, Rasch analysis evaluates unidimensionality – that the items measure a single, coherent latent trait (e.g., ideational praxis). Furthermore, I evaluated internal reliability of the tests – that any set of items within a test produces a similar measurement as the entire test. This study addresses the following research questions:

(1) What is the evidence for construct validity of data collected with each of the seven EASI praxis tests?
    a. Do the test items demonstrate uniformly positive point-measure correlations (i.e., do scores on each item correlate with overall test score?)
    b. Do 95% of items demonstrate adequate fit to the Rasch model?
    c. Do 90% of children demonstrate adequate fit to the Rasch model?
    d. Do the Rasch-generated step thresholds within rating scales progress in an orderly fashion?
    e. Does a Rasch principal components analysis (PCA) of standardized residuals reveal meaningful secondary dimensions in the data?

f. Does a differential item functioning analysis reveal invariance in item difficulty for male and female children?

g. Do the items form a logical hierarchy with sufficient item difficulty variation to match sample ability levels?

h. Do test-takers form a logical developmental hierarchy (i.e., do scores increase with increasing age?)

(2) What is the evidence for internal reliability of data collected using the 7 EASI praxis tests?

a. Do the data demonstrate adequate internal reliability based on the Rasch person reliability index?

b. Do the data reliably distinguish at least two levels of sensory integration based on the number of strata associated with the measure?

## Methods

In this study, I used Rasch analysis to evaluate evidence for validity and reliability of international normative data collected using the seven EASI praxis tests. I conducted separate analyses for each of the tests; this paper summarizes the results of these analyses.

### Participants

I drew data from the EASI International Normative Data Collection Project (maintained by the Collaboration for Leadership in Ayres Sensory Integration [CLASI]). The dataset comprises responses from 2563 children between the ages of 3 and 12 years. The dataset includes children from 51 countries. Inclusion criteria for normative data collection included: (1) chronological age between 3 years 0 months and 12 years 11 months; (2) typical development

(as reported by parents/primary caregivers); (3) no known medical, educational, mental health, or other developmental concerns. Exclusion criteria were: (1) known medical, educational, mental health, or other developmental concerns; (2) identified as having sensory integration concerns by OT, PT, or SLP; (3) receive(s/d) therapy services for learning disorders, ASD, ADHD, speech/language delays, regulatory issues, hypotonia, or DCD; (4) siblings who meet any of these exclusion criteria. Not all children completed every test; individual test sample sizes are reflected in Table 3.9. In Appendix C, I described the entire sample. The Rasch model is robust against missing data; therefore, I included children who did not complete all items within tests. I only omitted children who were missing more than 50% of test items.

**Procedure**

All normative data collectors completed an 8- to 10-hour online training course that covered EASI testing and scoring. At the conclusion of training, they completed a series of online scoring quizzes, achieving at least 80% accuracy against scores completed by a gold standard observer. Examiners gave the EASI in the child's native language, as appropriate. Examiners conducted EASI tests in locations convenient for children and families. This included clinics, children's homes, and research laboratories. Most examiners gave all 21 tests during a single 3- to 4-hour session; however, some required multiple sessions. Most common reasons for multiple sessions included scheduling conflicts or children's limited tolerance for extended testing. Examiners uploaded all data into a secure RedCap database managed by CLASI.

**Data Analysis**

I conducted all analyses using the Rasch-specific software program, Winsteps (Linacre, 2022). The Rasch model is a latent trait psychometric model that converts ordinal-level data (e.g., EASI raw scores) to interval-level measures (Bond, Yan & Heene, 2020). Both person

ability and item difficulty are estimated along the same log-odds unit ("logit") scale. Rasch is based upon two complementary assumptions; these assumptions can be expressed in terms of the EASI praxis tests, that: (1) easier items (i.e., items that require less developed praxis skills) are easier for all children, and (2) children with more well-developed praxis will successfully complete harder items compared with children with less well-developed praxis.

**Model Selection**

The Rasch model includes sub-models, including the dichotomous model (DM) and the rating scale model (RSM) (Bond, Yan & Heene, 2020). I employed DM to estimate item and person parameters for the two tests with only dichotomous items (PrFD and VPrC). For the four tests with trichotomous rating scales (0, 1, 2; PrP, PrS, PrOc, PrI, and VPrD), I used RSM. RSM specifies that all items within a test share a common rating scale, but may have polytomous rating scales (Wright, 1998). In addition to item difficulty and person ability estimates, RSM provides logit calibrations for rating scale categories. For PrI, tally items are scored on a 4-point rating scale while complexity and variety are each scored on trichotomous rating scales. For this test, I chose the grouped RSM in which items within a test share multiple common rating scales.

**Construct Validity**

In addition to generating item, person and rating scale calibrations, Winsteps provides several indicators that reveal the extent to which the data fit the Rasch model. I examined these indicators for evidence of construct validity: point-measure correlations, goodness-of-fit statistics, rating scale thresholds (for polytomous scales), item hierarchy, PCA of standardized residuals, and DIF.

***Point-measure Correlations.*** To determine if each item corresponds with the latent variable (i.e., that a higher score corresponds to improved sensory integration abilities), I

examined Pearson point-measure correlation coefficients between item and overall test measures. In the Rasch model, positive point-measure correlations suggest that items align with the construct (Bond, Yan & Heene, 2020). Of note, the magnitude of these correlations is less important than the directionality. To establish construct validity, all point-measure correlations should be positive.

***Goodness-of-fit Statistics (Items).*** I examined two kinds of mean-square goodness-of-fit statistics generated by Winsteps: infit and outfit. Infit statistics are "inlier-sensitive" or information-weighted to reduce the influence of off-target responses (i.e., people whose overall scores are far from the item measure). Outfit statistics are unweighted and typically reflect fit problems due to outliers. Mean-square measures show the amount of distortion of the measurement system. Ideal mean-square value is 1.0. Acceptable mean-square fit statistics vary based on sample size and the consequences of misfitting items. For this study, given the relatively large sample size, I accepted relatively conservative values between 0.7 and 1.3 as evidence of fit to the Rasch model (Linacre, 2002).

To demonstrate sufficient evidence for construct validity, at least 95% of items on each test should show adequate fit to the Rasch model. For tests with fewer than 20 items, a single misfitting item would fall below this threshold. Because I might expect at least one item to fail to fit due to chance alone, I expected *either* 95% of items to fit the model, or *no more than one* item to fail to fit.

***Goodness-of-fit Statistics (Children).*** Person fit statistics are calculated and interpreted in the same way as item fit statistics. As a rule, people behave less predictably than items (Bond, Yan, & Heene, 2020). Further, given that I had relatively few items compared to people, I selected less stringent criteria for acceptable mean-squares (0.5 to 1.5). Individuals who overfit

the model (i.e., behave too predictably, mean-square < 0.5) are unlikely to distort or degrade the measurement system (Linacre, 2015). Therefore, I only considered children's data as failing to fit if they underfit the model (MnSq > 1.5). As evidence of strong construct validity, I expected that data from 90% of children would fit the model for each test

***Rating Scale Analysis.*** For all rating scales, I examined (1) rating scale goodness-of-fit statistics. Mean-square between 0.7 and 1.3 suggest fit to the Rasch model (Linacre, 2021); and (2) Observed average person measure associated with each category. The observed average person measure associated with each category should demonstrate orderly progression: the lowest category should correspond with the lowest average person measure (Bond, Yan & Heene, 2020). Finally, for tests with polytomous rating scales (all except PrFD and VPrC), I examined (3) Andrich thresholds (i.e., the person ability measure at which a person is equally likely to use two adjacent categories). Andrich thresholds should progress in an orderly fashion, such that the lowest step threshold corresponds to the threshold between the two lowest categories and so forth (Bond, Yan & Heene, 2020; Linacre, 2018). Thresholds are not calculated for dichotomous scales.

For PrI tally items, I constructed a rating scale based upon the number of actions demonstrated by the child. I tried several rating scales and, based on the criteria described above, selected the rating scale that best fit the data (described in Table 3.10).

***Item Hierarchy.*** I assessed the item hierarchies in two ways. First, I compared the mean item measure and the mean person measure. In the Rasch model, the mean item measure is set at 0.0 logits. Mean person measure close to 0.0 indicates a match between the sample's sensory integration ability and the scale difficulty (Bond, Yan & Heene, 2020). Second, I visually inspected the Winsteps-generated Wright maps. The Wright map provides a hierarchy of items

and persons along a logit scale, ranging from lowest to higher measures. These items should be ordered logically so that theoretically-more-difficult items are associated with higher item difficulty measures. The most robust way to examine the logic of an item hierarchy is to compare the items with existing literature; however, few previous studies have examined praxis items in this level of detail. Therefore, I visually examined the item hierarchies to ensure that they matched theoretical expectations. I also examined the Wright maps to evaluate the spread of items; large gaps in item difficulty indicate a need for more items, while items grouped together suggest redundancy.

***Person Hierarchy.*** I assessed person hierarchies in two ways. The Winsteps-generated person maps show child scores along a continuous, interval-level scale on the right side of the figure and items along the left side (see, for example, Figure 3.1). The interval scale is broken into .2 logit levels. I averaged the ages of children on each level and visually inspected the map for evidence that average age increased with increasing scores. Second, I evaluated the strength of this relationship by conducting bivariate Pearson correlations between Rasch-generated child measure scores and children's age in months. Given the developmental nature of sensory integration constructs, I expected at least moderate correlation coefficients ($\geq$.30; Cohen, 1988). I confirmed normality of all variables (age in months and EASI measure scores) using the methods described by Kim (2013) for large sample sizes ($N > 300$).

***Principal Components Analysis.*** While goodness-of-fit statistics and other evidence described above examine the extent to which the construct is unidimensional, PCA provides evidence of the strength of additional dimensions in the data (i.e., multidimensionality). PCA deconstructs model residuals to identify additional dimensions in the data (i.e., item response patterns not explained by the Rasch model). Eigenvalues estimate the strength of these

dimensions (called contrasts). I considered contrasts to be strong enough to refute unidimensionality of the construct if the following conditions were met (Linacre, 2018): (1) There are contrasts with eigenvalues > 2 (i.e., with the strength of more than 2 items); (2) Item subsets within contrasts demonstrate disattenuated correlations < 0.57, indicating that item subsets likely measure different latent variables. According to Wright and Stone (1979), unidimensionality is essential to good construct measurement; evidence of multiple dimensions suggests that the items should be scored as multiple, separate instruments.

***Measurement Invariance.*** I used Rasch DIF analyses to examine the measurement invariance of the praxis tests based on sex (i.e., that test items are not biased based on the child's sex). Using the Rasch-Welch DIF method (Linacre, 2020), I compared item difficulty estimates for males and females. DIF contrasts (i.e., the difference between difficulty estimates) for males and females should be no larger than .43 logits (Zwick, Thayer & Lewis, 1999) to be considered negligible. I also conducted *t*-tests of item difficulty to examine the likelihood that DIF could be caused by chance alone. Items with both DIF contrast > .43 *and* $p \leq .05$ should be considered problematic and should be removed or targeted for revision. Given the large sample size in this study, I did not consider items to show bias if contrasts were significant but smaller than .43 logits.

**Internal Reliability**

I evaluated internal reliability based on two Winsteps-generated indices. The first, person reliability index, is the Rasch equivalent to Cronbach's alpha and represents the amount of variance that can be reproduced by the Rasch model (Wright & Masters, 1982). A person reliability index greater than 0.80 suggests strong evidence for internal reliability; greater than .70 is adequate (Bond, Yan & Heene, 2020).

The strata value is an additional measure of reliability that represents the number of levels of ability that the measure can distinguish (Wright & Masters, 1982). Winsteps generates a separation index (G) that I converted to strata using the formula:

$$Strata = \frac{4G+1}{3}.$$

Strata should be at least 2.0 to establish evidence for sufficient internal reliability (Bond, Yan & Heene, 2020). Given the developmental nature of praxis and the large age range of our sample, I expect higher strata values (i.e., I expect more levels of ability to be represented by the items). Therefore, I will consider strata values acceptable at 2.0 and strong at 3.0 or more.

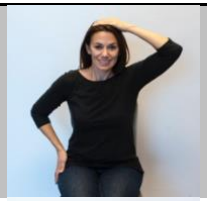
## Results

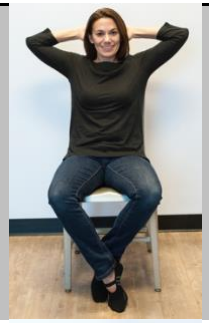
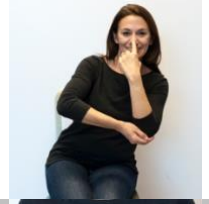
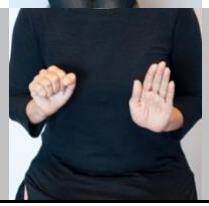
### Construct Validity

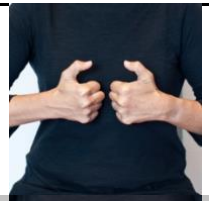
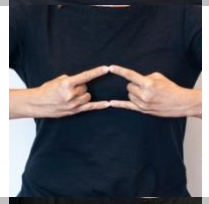
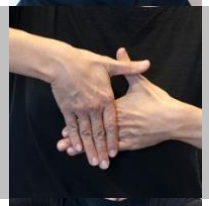
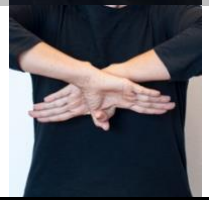
Tables 3.2-3.8 contain item measures, point-measure correlations, and fit statistics for each of the praxis tests. All items showed positive point-measure correlations. For PrP, two items (Face 1 and Face 4) demonstrated underfit; 91.7% of items (22/24) fit the model. For VPrD, five items failed to fit (Designs 1, 2, 5, 6 and 7); 79.2% of items (19/24) fit the model. For VPrC, two items (Left Chair Position 2 and Right Chair Position 2) failed to fit; 94.1% of items (32/34) fit the model. For PrS, PrOc, PrI, and PrFD, 100% of items fit the model.

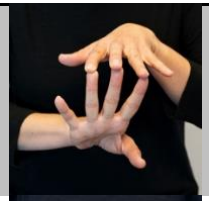
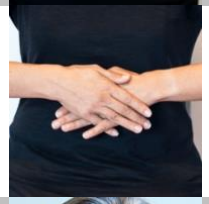
**Table 3.2**

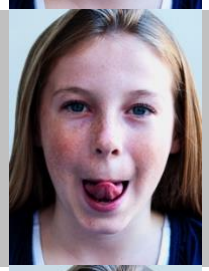
*Praxis: Positions Item Measures and Fit Statistics and DIF Statistics*

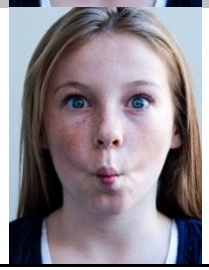
| | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
|  | PRPB 1 | -0.77 (0.05) | 1.06 | 0.95 | 0.45 | .11 |
|  | PRPB 2 | -1.22 (0.05) | 0.91 | 1.04 | 0.38 | .00 |
|  | PRPB 3 | -0.30 (0.04) | 0.78 | 0.92 | 0.48 | -.07 |
|  | PRPB 4 | -0.36 (0.04) | 0.86 | 1.02 | 0.44 | -.14 |

| | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
|  | PRPB 5 | -0.69 (0.04) | 1.03 | 1.05 | 0.44 | -.02 |
|  | PRPB 6 | -0.09 (0.04) | 0.89 | 0.85 | 0.54 | .18* |
|  | PRPB 7 | 0.58 (0.03) | 0.97 | 0.96 | 0.55 | -.06 |
|  | PRPB 8 | 0.24 (0.04) | 0.86 | 0.90 | 0.55 | .00 |
|  | PRPH 1 | -1.02 (0.05) | 0.96 | 1.04 | 0.39 | -.25* |

| | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
|  | PRPH 2 | -0.69 (0.04) | 0.95 | 1.06 | 0.42 | -.17 |
|  | PRPH 3 | 0.68 (0.03) | 0.83 | 0.89 | 0.55 | -.12 |
|  | PRPH 4 | 0.70 (0.03) | 0.89 | 0.83 | 0.62 | -.38* |
|  | PRPH 5 | 1.21 (0.03) | 0.84 | 0.87 | 0.6 | -.12 |
|  | PRPH 6 | 2.16 (0.03) | 1.02 | 0.99 | 0.61 | -.02 |

| | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
|  | PRPH 7 | 1.43 (0.03) | 0.81 | 0.79 | 0.65 | .07 |
|  | PRPH 8 | 1.94 (0.03) | 0.93 | 0.91 | 0.62 | .00 |
|  | **PRPF 1** | -0.55 (0.04) | **1.50** | **1.46** | 0.39 | -.35* |
|  | PRPF 2 | -1.96 (0.07) | 1.12 | 1.02 | 0.32 | .31* |

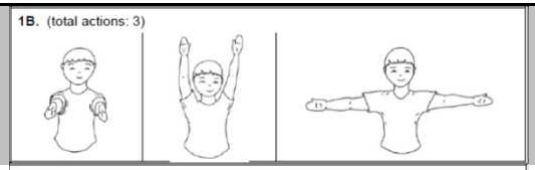
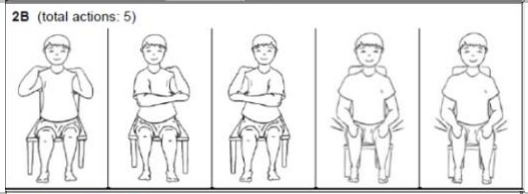
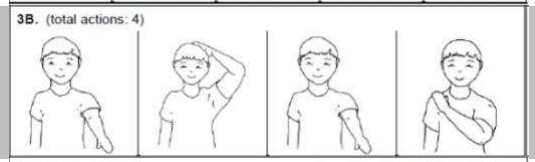
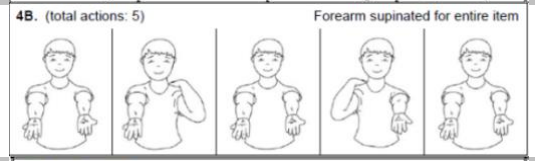
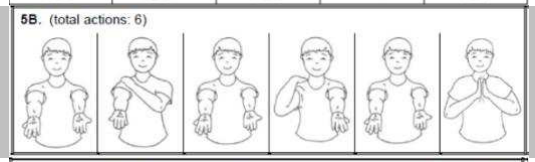
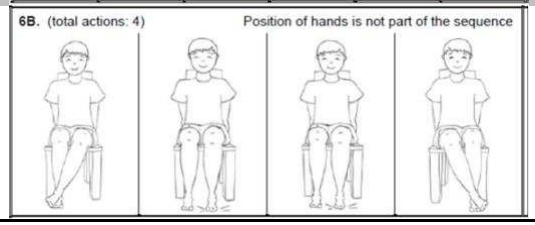
| | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
|  | PRPF 3 | 0.65 (0.03) | 1.27 | 1.21 | 0.53 | .20* |
|  | **PRPF 4** | -0.51 (0.04) | **1.36** | 1.21 | 0.41 | .21* |
|  | PRPF 5 | -0.45 (0.04) | 1.24 | 1.16 | 0.41 | .15 |
|  | PRPF 6 | -0.08 (0.04) | 1.30 | 1.23 | 0.46 | .14 |

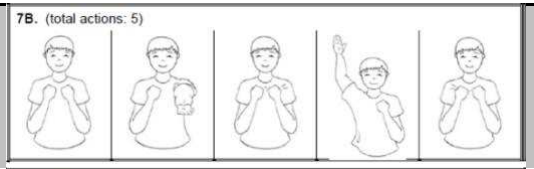
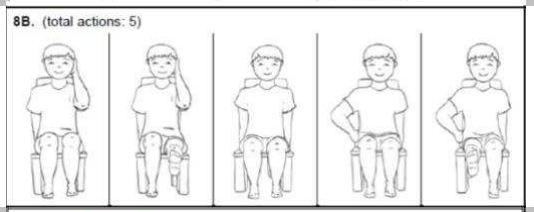
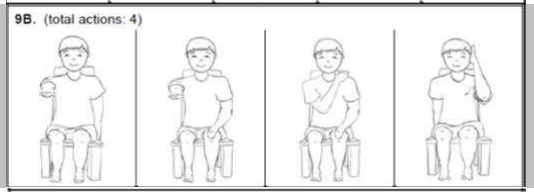
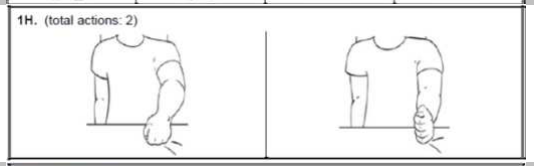
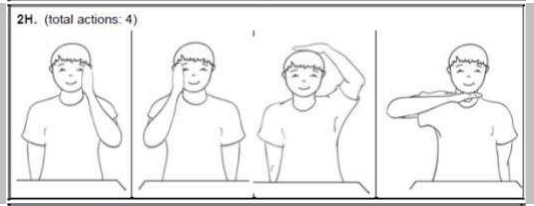
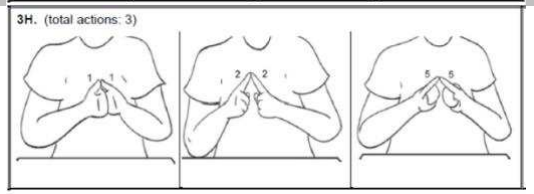
| | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
|  | PRPF 7 | -0.97 (0.05) | 1.10 | 0.98 | 0.42 | .41* |
|  | PRPF 8 | 0.06 (0.04) | 1.09 | 1.08 | 0.49 | .24* |

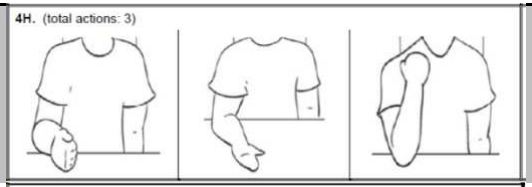
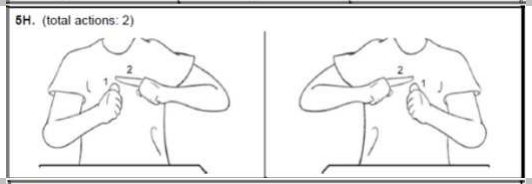
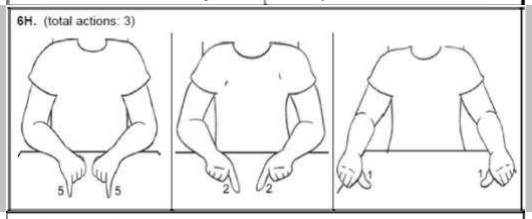
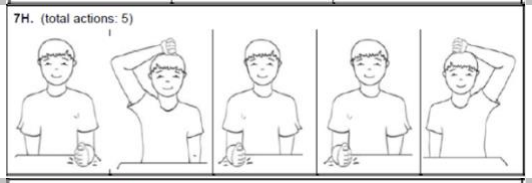
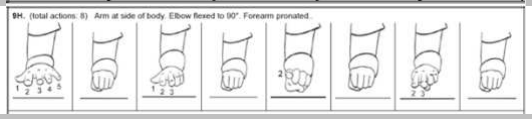
*Note.* **Bold** indicates misfitting items, * indicates significance ($p < .05$); PRPB = Praxis Positions Body; PRPH = Praxis Positions Hands; PRPF = Praxis Positions Face

**Table 3.3**

*Praxis: Sequences Item Measures and Fit Statistics and DIF Statistics*

| | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
|  | PRSB 1 | -0.75 (0.04) | 1.07 | 1.07 | 0.46 | -.04 |
|  | PRSB 2 | 0.36 (0.03) | 0.97 | 1.02 | 0.55 | .09 |
|  | PRSB 3 | -0.01 (0.03) | 1.01 | 0.99 | 0.55 | -.06 |
|  | PRSB 4 | -0.24 (0.03) | 0.99 | 0.92 | 0.6 | .00 |
|  | PRSB 5 | 0.77 (0.03) | 0.93 | 0.93 | 0.6 | .06 |
|  | PRSB 6 | 0.16 (0.03) | 0.93 | 0.93 | 0.55 | -.12 |

| | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
|   | PRSB 7 | 0.66 (0.03) | 0.92 | 0.89 | 0.6 | .00 |
|   | PRSB 8 | 0.96 (0.03) | 0.97 | 0.99 | 0.56 | .06 |
|   | PRSB 9 | 1.69 (0.03) | 0.91 | 0.96 | 0.52 | -.07 |
|   | PRSH 1 | -1.84 (0.05) | 1.19 | 1.12 | 0.37 | -.08 |
|   | PRSH 2 | -0.02 (0.03) | 0.95 | 0.99 | 0.5 | -.14* |
|   | PRSH 3 | -0.36 (0.04) | 1.1 | 1.02 | 0.52 | .11 |

| | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
|  | PRSH 4 | -0.53 (0.04) | 1.08 | 1.07 | 0.49 | .08 |
|  | PRSH 5 | -0.06 (0.03) | 1.2 | 1.11 | 0.56 | .00 |
|  | PRSH 6 | -0.08 (0.03) | 1.07 | 1.03 | 0.55 | -.07 |
|  | PRSH 7 | 0.75 (0.03) | 0.95 | 0.93 | 0.56 | -.10 |
|  | PRSH 8 | 0.77 (0.03) | 0.92 | 0.91 | 0.56 | .12 |
|  | PRSH 9 | 2.00 (0.03) | 0.94 | 0.89 | 0.55 | .08 |
| With mouth closed, sniff air in through the nose, three times | PRSF 1 | -1.84 (0.05) | 1.21 | 1.00 | 0.42 | .08 |
| Make non-voiced sounds P-T-P-K | PRSF 2 | -0.52 (0.04) | 1.01 | 0.97 | 0.53 | .00 |

| | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
| Exaggerate sucking air in through mouth; blow out; suck air in | PRSF 3 | -1.2 (0.04) | 1.18 | 1.17 | 0.45 | .00 |
| Blink two times; pucker lips | PRSF 4 | -1.09 (0.04) | 1.23 | 1.22 | 0.44 | .14 |
| Make non-voiced sounds P-T-K; click teeth together, two times | PRSF 5 | -0.29 (0.03) | 1.06 | 1.06 | 0.52 | -.12 |
| Open mouth; touch tongue to top lip; touch tongue to bottom lip, two times | PRSF 6 | -0.02 (0.03) | 1 | 0.97 | 0.55 | .00 |
| Raise eyebrows two times; blink both eyes; raise eyebrows | PRSF 7 | 0.16 (0.03) | 1.01 | 0.99 | 0.56 | -.05 |
| Make non-voiced sounds P-K-P-T-K | PRSF 8 | 0.61 (0.03) | 0.97 | 0.97 | 0.54 | -.07 |
| Put air in R cheek, put air in L cheek; put air in both cheeks; quickly pull lips apart (with force to make a loud sound) | PRSF 9 | -0.03 (0.03) | 0.98 | 0.99 | 0.55 | -.05 |

*Note.* * indicates significance ($p < .05$); PRSB = Praxis Sequences Body; PRSH = Praxis Sequences Hands; PRSF = Praxis Sequences Face

**Table 3.4**

*Praxis: Following Directions Item Measures and Fit Statistics and DIF Statistics*

| Item Description | Name | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
| Put both arms up | PRFDB 1 | -3.18 (0.16) | 1.02 | 1.17 | 0.3 | -.30 |
| Bend to one side | PRFDB 2 | -0.14 (0.07) | 1.09 | 1.11 | 0.49 | .24 |
| Lift your foot and put it out to the front | PRFDB 3 | -0.91 (0.08) | 1.12 | 1.21 | 0.42 | .08 |
| Put one hand forward and one foot back | PRFDB 4 | -0.28 (0.07) | 0.91 | 0.9 | 0.54 | .17 |
| Turn your head to the side, cross your legs, and cross your arms | PRFDB 5 | 0.64 (0.06) | 0.95 | 0.91 | 0.58 | .02 |
| Put one arm up, one arm down and one foot back | PRFDB 6 | 0.11 (0.06) | 0.92 | 0.87 | 0.56 | -.06 |
| Put your thumb on one finger of the other hand | PRFDH 1 | 0.03 (0.06) | 0.96 | 0.96 | 0.54 | -.10 |
| Make a ball with this piece of paper | PRFDH 2 | 0.29 (0.06) | 1.1 | 1.1 | 0.5 | .00 |

| Item Description | Name | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
| (give the paper to the child) | | | | | | |
| Make one hand look like a ball and cover it with the other hand | PRFDH 3 | 0.09 (0.06) | 0.9 | 0.8 | 0.57 | -.05 |
| Hide your thumb with your fingers | PRFDH 4 | -0.2 (0.07) | 0.92 | 0.81 | 0.54 | .05 |
| Touch a thumb to two fingers of the other hand | PRFDH 5 | 0.34 (0.06) | 0.84 | 0.77 | 0.61 | .10 |
| Hold your other hand and move it in a circle 3 times | PRFDH 6 | 2.34 (0.05) | 1.09 | 1.03 | 0.58 | -.10 |
| Open your mouth and cover your teeth with your lips | PRFDF 1 | 1.22 (0.05) | 1.07 | 1.08 | 0.56 | -.14 |
| Open your eyes really big and close your mouth all the way | PRFDF 2 | -0.5 (0.07) | 1.11 | 1.15 | 0.45 | -.13 |

| Item Description | Name | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
| Move your mouth from side to side | PRFDF 3 | 0.46 (0.06) | 1.07 | 1.08 | 0.52 | .41* |
| Cover your top lip with your tongue | PRFDF 4 | 0.87 (0.06) | 1.06 | 1.06 | 0.54 | -.27* |
| Raise your eyebrows three times | PRFDF 5 | 0.73 (0.06) | 0.94 | 0.9 | 0.59 | .00 |
| Close your eyes and open your mouth | PRFDF 6 | -1.91 (0.1) | 1.01 | 1.07 | 0.37 | .11 |

*Note.* * indicates significance ($p < .05$); PRFDB = Praxis Following Directions Body; PRFDH = Praxis Following Directions Hands; PRFDF = Praxis Following Directions Face

**Table 3.5**

*Ocular Praxis Item Measures and Fit Statistics and DIF Statistics*

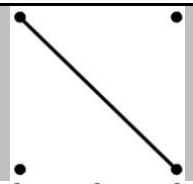
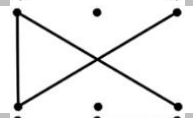
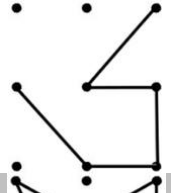
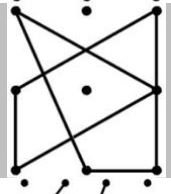
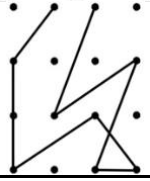
| Item Description | Item Name | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
| Move eyes looking upward, hold two seconds, return to neutral | PROC 1 | -1.22 (0.05) | 1.19 | 1.23 | 0.63 | -.06 |
| Move eyes looking downward, hold two seconds, return to neutral | PROC 2 | -1.29 (0.05) | 1.15 | 1.21 | 0.63 | .32* |
| Move eyes looking upper left, hold two seconds, return to neutral | PROC 3 | -0.43 (0.05) | 1.21 | 1.18 | 0.68 | -.22* |
| Move eyes looking upper right to upper left, two times, return to neutral | PROC 4 | 0.64 (0.04) | 0.97 | 0.96 | 0.76 | -.21* |
| Move eyes upper left to lower right, two times, return to neutral | PROC 5 | 0.65 (0.04) | 0.79 | 0.75 | 0.8 | .00 |
| Move eyes lower left to upper | PROC 6 | 0.92 (0.04) | 0.86 | 0.83 | 0.79 | .17* |

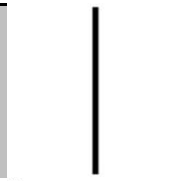
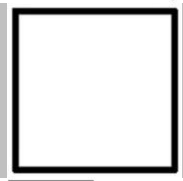
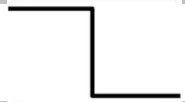
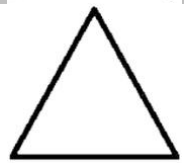
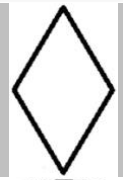
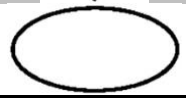
| Item Description | Item Name | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
| right, two times, return to neutral | | | | | | |
| Move eyes horizontal right to left, left to right, two times, return to neutral | PROC 7 | 0.13 (0.04) | 1.01 | 0.99 | 0.74 | -.06 |
| Move eyes straight up, then straight down, two times, return to neutral | PROC 8 | 0.58 (0.04) | 0.99 | 0.97 | 0.76 | .05 |

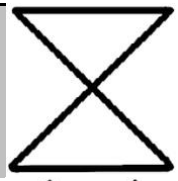
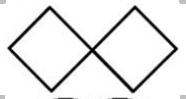
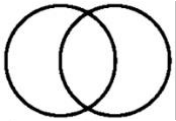
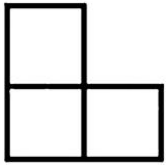
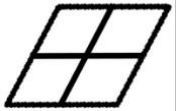
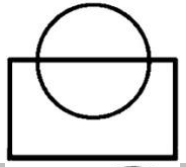
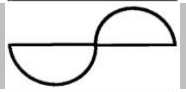
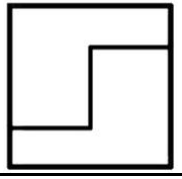
*Note.* * indicates significance ($p < .05$)

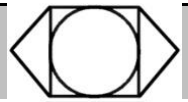
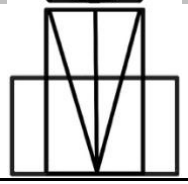
**Table 3.6**

*Visual Praxis: Designs Item Measures and Fit Statistics and DIF Statistics*

| | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
|  | **VPRD 1** | -2.67 (0.07) | **1.67** | **2.35** | 0.6 | .00 |
|  | **VPRD 2** | -0.56 (0.05) | **1.38** | **1.47** | 0.7 | .13 |
|  | VPRD 3 | -0.65 (0.05) | 1.04 | 0.95 | 0.73 | -.10 |
|  | VPRD 4 | -0.53 (0.05) | 1.06 | 1.08 | 0.71 | .00 |
|  | **VPRD 5** | 1.06 (0.04) | **1.37** | 1.23 | 0.69 | -.21* |
|  | **VPRD 6** | 1.24 (0.04) | **1.39** | 1.24 | 0.68 | -.33* |

| | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
|  | **VPRD 7** | -4.52 (0.1) | **1.71** | **3.79** | 0.45 | .24 |
|  | VPRD 8 | -2.84 (0.07) | 1.17 | 1.28 | 0.63 | .21 |
|  | VPRD 9 | -0.68 (0.05) | 1.02 | 1.04 | 0.71 | .16 |
|  | VPRD 10 | -1.13 (0.05) | 1.16 | 1.26 | 0.68 | -.08 |
|  | VPRD 11 | -0.86 (0.05) | 0.98 | 1.00 | 0.71 | .09 |
|  | VPRD 12 | -0.64 (0.05) | 0.88 | 0.91 | 0.73 | .15 |
|  | VPRD 13 | 0.48 (0.04) | 0.76 | 0.70 | 0.77 | .00 |
|  | VPRD 14 | 0.05 (0.04) | 0.94 | 1.20 | 0.71 | .14 |

| | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
|  | VPRD 15 | 0.2 (0.04) | 0.86 | 0.91 | 0.74 | .12 |
|  | VPRD 16 | 1.24 (0.04) | 0.82 | 0.81 | 0.75 | .00 |
|  | VPRD 17 | 0.44 (0.04) | 0.82 | 0.87 | 0.73 | .08 |
|  | VPRD 18 | 1.18 (0.04) | 0.91 | 0.93 | 0.72 | .00 |
|  | VPRD 19 | 1.98 (0.04) | 0.92 | 0.83 | 0.73 | -.06 |
|  | VPRD 20 | 0.54 (0.04) | 0.79 | 0.82 | 0.74 | .13 |
|  | VPRD 21 | 1.16 (0.04) | 0.87 | 0.82 | 0.74 | .00 |
|  | VPRD 22 | 0.95 (0.04) | 0.72 | 0.72 | 0.76 | -.17* |

| | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
|  | VPRD 23 | 2.31 (0.04) | 0.84 | 0.81 | 0.73 | -.07 |
|  | VPRD 24 | 2.21 (0.04) | 0.92 | 0.89 | 0.72 | .00 |

*Note*. **Bold** indicates misfitting items, * indicates significance ($p < .05$)

**Table 3.7**

*Visual Praxis: Construction Item Measures and Fit Statistics and DIF Statistics*

| Visual Model for the Child to Recreate | Item Name | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
|  | Left Chair Position 1 | -1.31 (0.07) | 0.97 | 0.81 | 0.42 | -.05 |
| | Left Chair Position 2 | -0.96 (0.07) | 1.17 | 1.58 | 0.32 | .04 |
| | Left Chair Orientation 1 | -0.88 (0.07) | 1.01 | 0.9 | 0.43 | -.16 |
| | Left Chair Orientation 2 | 0.44 (0.05) | 0.98 | 0.93 | 0.5 | .00 |
| | Left Chair Orientation 3 | 0.73 (0.05) | 0.98 | 0.97 | 0.52 | -.06 |
| | Left Chair Orientation 4 | 0.81 (0.05) | 0.92 | 0.84 | 0.55 | .00 |
| | Right Chair Position 1 | -1.35 (0.08) | 0.97 | 1.01 | 0.41 | -.30* |
| | Right Chair Position 2 | 2.23 (0.05) | 1.26 | 1.56 | 0.41 | -.15 |
| | Right Chair Orientation 1 | -0.47 (0.06) | 0.99 | 0.88 | 0.46 | .07 |
| | Right Chair Orientation 2 | 0.31 (0.05) | 0.97 | 0.91 | 0.51 | -.21* |
| | Right Chair Orientation 3 | 0.68 (0.05) | 0.92 | 0.87 | 0.54 | -.12 |
| | Right Chair Orientation 4 | 0.73 (0.05) | 0.94 | 0.87 | 0.54 | -.09 |
| | Yoga Mat Position 1 | 0.66 (0.05) | 1.11 | 1.17 | 0.45 | .07 |
| | Yoga Mat Position 2 | -0.48 (0.06) | 1 | 1.03 | 0.44 | .33* |
| | Yoga Mat Position 3 | 0.93 (0.05) | 1.16 | 1.2 | 0.43 | .00 |
| | Yoga Mat Orientation 1 | 0.17 (0.05) | 0.89 | 0.82 | 0.54 | .11 |
| | Yoga Mat Orientation 2 | 0.39 (0.05) | 0.91 | 0.83 | 0.54 | .19 |
| | Rice Bottle Position 1 | -0.47 (0.06) | 0.97 | 0.98 | 0.46 | -.12 |
| | Rice Bottle Position 2 | 0.51 (0.05) | 1.07 | 1.1 | 0.46 | -.09 |
| | Rice Bottle Position 3 | -0.85 (0.07) | 0.84 | 0.69 | 0.51 | .30* |
| | Rice Bottle Orientation 1 | -0.9 (0.07) | 0.94 | 0.82 | 0.46 | .21 |
| | Rice Bottle Orientation 2 | 0.4 (0.05) | 0.91 | 0.84 | 0.54 | .00 |
| | Rice Bottle Orientation 3 | 0.01 (0.05) | 1 | 1.01 | 0.47 | .00 |
| | Lucite Stand Position 1 | -0.33 (0.06) | 1.1 | 1.22 | 0.4 | -.02 |
| | Lucite Stand Position 2 | 0.72 (0.05) | 1.23 | 1.37 | 0.39 | -.15 |
| | Lucite Stand Orientation 1 | -1.62 (0.08) | 0.9 | 0.66 | 0.44 | .00 |

| Visual Model for the Child to Recreate | Item Name | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
| | Lucite Stand Orientation 2 | -0.45 (0.06) | 1 | 0.94 | 0.45 | .11 |
| | Cardstock Strip Position 1 | -0.55 (0.06) | 0.93 | 0.92 | 0.48 | .04 |
| | Cardstock Strip Orientation 1 | 0.25 (0.05) | 0.91 | 0.92 | 0.52 | .09 |
| | Cardstock Strip Orientation 2 | 0.65 (0.05) | 0.96 | 0.92 | 0.53 | .10 |

*Note.* **Bold** indicates misfitting items, * indicates significance ($p < .05$)

**Table 3.8**

*Praxis: Ideation Item Measures and Fit Statistics and DIF Statistics*

| Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|
| Chair Tally | 0.51 (0.03) | 1.05 | 1.05 | 0.70 | -.10 |
| Chair Variety | 0.22 (0.04) | 0.97 | 0.94 | 0.68 | -.18* |
| Chair Complexity | 0.97 (0.04) | 0.95 | 0.89 | 0.66 | -.12 |
| Hands Tally | -0.59 (0.03) | 1.13 | 1.11 | 0.71 | .11 |
| Hands Variety | -0.55 (0.04) | 0.92 | 0.90 | 0.70 | .13 |
| Hands Complexity | 0.15 (0.04) | 0.97 | 0.93 | 0.68 | .06 |
| Objects Tally | 0.48 (0.03) | 1.14 | 1.17 | 0.64 | -.15* |
| Objects Variety | 0.1 (0.04) | 0.97 | 1.01 | 0.67 | .00 |
| Objects Complexity | 0.43 (0.04) | 0.97 | 0.96 | 0.66 | .06 |
| Body Tally | -0.74 (0.03) | 1.14 | 1.11 | 0.71 | .00 |
| Body Variety | -0.83 (0.04) | 0.89 | 0.89 | 0.70 | .10 |
| Body Complexity | -0.15 (0.04) | 0.87 | 0.86 | 0.70 | .11 |

*Note.* * indicates significance ($p < .05$)

Table 3.9 contains the results of person fit analysis. Five tests fell below our desired threshold (90% person fit): PrP, PrS, PrOc, VPrD and PrI. Two tests met our criteria: PrFD and VPrC. Notably, the tests with more children who did not fit still had at least 84.8% person fit.

**Table 3.9**

*Person Fit Analysis*

| Test | Children with Misfitting Data | Total Number of Children | % Fitting Children |
|---|---|---|---|
| Praxis: Positions | 351 | 2489 | 85.9% |
| Praxis: Sequences | 277 | 2438 | 88.6% |
| Praxis: Following Directions | 205 | 2421 | 92.5% |
| Ocular Praxis | 368 | 2428 | 84.8% |
| Visual Praxis: Designs | 369 | 2146 | 85.3% |
| Visual Praxis: Construction | 172 | 2350 | 92.7% |
| Praxis: Ideation | 254 | 2441 | 85.5% |

Table 3.10 contains the results of rating scale analyses. Across all tests, fit statistics and rating scale categories were used more than 10% of the time. Furthermore, all polytomous rating scales demonstrated acceptable Andrich thresholds. Finally, observed average person measures increased monotonically for each rating scale category.

**Table 3.10**

*Rating Scale Analysis for Praxis Tests*

| Item Type | Rating Scale Category | % Used | Infit MnSq | Outfit MnSq | Andrich Threshold[1] | Observed Average |
|---|---|---|---|---|---|---|
| | | | Praxis: Positions | | | |
| Accuracy | 0 | 12.73% | 1.03 | 1.14 | – | -0.46 |
| | 1 | 23.22% | 0.97 | 0.94 | -0.46 | .77 |
| | 2 | 64.06% | 1.00 | 1.00 | 0.46 | 2.11 |
| | | | Praxis: Sequences | | | |
| Accuracy | 0 | 19.87% | 1.04 | 1.12 | – | -0.60 |
| | 1 | 27.01% | .95 | .87 | -.36 | 0.47 |
| | 2 | 53.12% | 1.00 | 1.01 | .36 | 1.59 |
| | | | Praxis: Following Directions | | | |
| Accuracy | 0 | 20.73% | 1.00 | 1.00 | – | -0.03 |
| | 1 | 79.27% | 1.01 | .99 | – | 2.18 |

| Item Type | Rating Scale Category | % Used | Infit MnSq | Outfit MnSq | Andrich Threshold[1] | Observed Average |
|---|---|---|---|---|---|---|
| | | | Ocular Praxis | | | |
| Accuracy | 0 | 15.30% | .97 | .99 | – | -0.98 |
| | 1 | 20.38% | .98 | 1.02 | -.62 | 0.58 |
| | 2 | 64.32% | 1.03 | 1.02 | .62 | 1.92 |
| | | | Visual Praxis: Designs | | | |
| Accuracy | 0 | 18.27% | .99 | 1.41 | – | -2.08 |
| | 1 | 29.70% | .97 | 1.14 | -1.15 | 0.53 |
| | 2 | 52.03% | 1.01 | 1.02 | 1.15 | 2.95 |
| | | | Visual Praxis: Construction | | | |
| Accuracy | 0 | 27.22% | 0.99 | 0.95 | – | 0.04 |
| | 1 | 72.78% | 1.02 | 1.07 | – | 1.77 |
| | | | Praxis: Ideation | | | |
| Tally | 0 | 16.35% | 1.14 | 1.11 | – | -1.82 |
| | 1 | 37.95% | 1.09 | 1.08 | -2.20 | -.46 |
| | 2 | 27.15% | 1.11 | 1.12 | 0.48 | 0.70 |
| | 3 | 18.55% | 1.13 | 1.14 | 1.73 | 1.82 |
| Variety | 0 | 26.89% | 0.93 | 0.90 | – | -1.33 |
| | 1 | 38.67% | 0.94 | 0.92 | -0.98 | 0.18 |
| | 2 | 34.44% | 0.95 | 0.98 | 0.98 | 1.55 |
| Complexity | 0 | 36.30% | 0.87 | 0.87 | – | -1.69 |
| | 1 | 38.99% | 0.91 | 0.82 | -0.98 | 0.06 |
| | 2 | 24.71% | 1.02 | 1.03 | 0.98 | 1.11 |

[1]Andrich Thresholds only calculated for non-dichotomous scales

PCA (Principal Components Analysis) supported unidimensionality of each test (see Table 3.11). Eigenvalues of the largest contrasts for PrP, PrS, PrFD and PrOc fell below 2.0, suggesting a single dimension for these tests. VPrD, VPrC and PrI showed eigenvalues greater than 2.0; however, for each of these tests, the disattenuated correlations between the item subsets were greater than .57 ($r$ = .92, .93 and .60, respectively). For all tests, the Rasch dimension explained a much greater proportion of the variance compared to the largest contrast detected by PCA.

**Table 3.11**

Principal Components Analysis of Standardized Rasch Residuals

| Test | Eigenvalue of Largest Contrast | Variance Explained by Largest Contrast | Variance Explained by Rasch Dimension |
|---|---|---|---|
| Praxis: Positions | 1.61 | 3.8% | 44.0% |
| Praxis: Sequences | 1.92 | 4.1% | 42.2% |
| Praxis: Following Directions | 1.39 | 5.2% | 33.0% |
| Ocular Praxis | 1.45 | 9.1% | 49.6% |
| Visual Praxis: Designs | 2.19 | 3.5% | 61.7% |
| Visual Praxis: Construction | 2.06 | 4.9% | 28.1% |
| Praxis: Ideation | 2.94 | 12.5% | 48.9% |

DIF analysis revealed strong evidence for measurement invariance between male and female children. Tables 3.2-3.8 contain DIF contrasts for each item. While each test showed some significant contrasts, no contrasts exceeded .43 logits (the unit of measure in the Rasch model, which generally ranges from about +3 to – 3 logits). Therefore, all items showed no/negligible DIF based on sex.

Figures 3.1-3.7 show the Winsteps-generated Wright maps for each of the praxis tests. Children were clustered at the top of the scales for PrFD, PrOc and VPrC, suggesting that items are too easy for the children in the sample. Items and children followed an approximately symmetric distribution for the remaining tests. Notably, the mean and standard deviations shown on these figures omit children with maximum and minimum scores, as these children's measures cannot be estimated in the Rasch model. Table 3.12 contains the mean measure scores including *all* participating children.

Upon inspection, the Wright maps presented here (Figures 3.1-3.7) presented logical item hierarchies. For PrP, PrS, and PrFD, items with more complex requirements (i.e., intricate hand items) had harder measure scores than simple, full-body items. Similarly, for PrOc, multi-step eye movements and those involving diagonal motion were more difficult than simple vertical or horizontal eye motions. Similarly, on VPrD, simple designs (e.g., a vertical line, Item 7) fell lower on the hierarchy than complex designs (e.g., Items 23 and 24). Further, designs with dot

grids were overall easier than free-handed designs. For VPrC, detailed orientation and position items (e.g., that items are touching rather than separated) were more difficult than simple items such as a chair placed on each side of the table. For PrI, items involving external objects were more difficult than those involving just the children's hands/bodies.

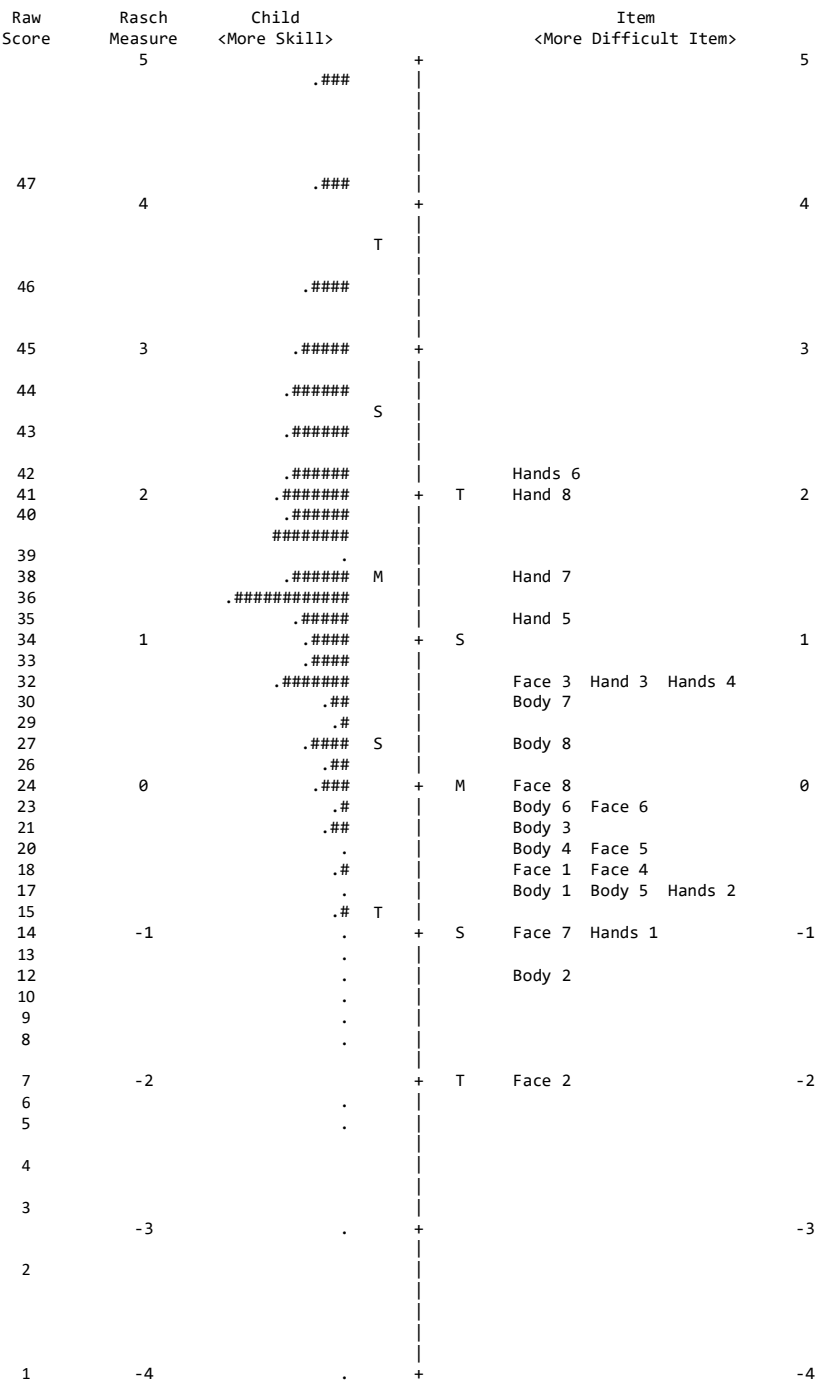



**Figure 3.1**

*Praxis: Positions Wright Map*

*Note*. # = 20 children, . = 1 to 19 children, | = latent trait (praxis skill), M = item/child mean, S = 1 SD, T = 2 SD

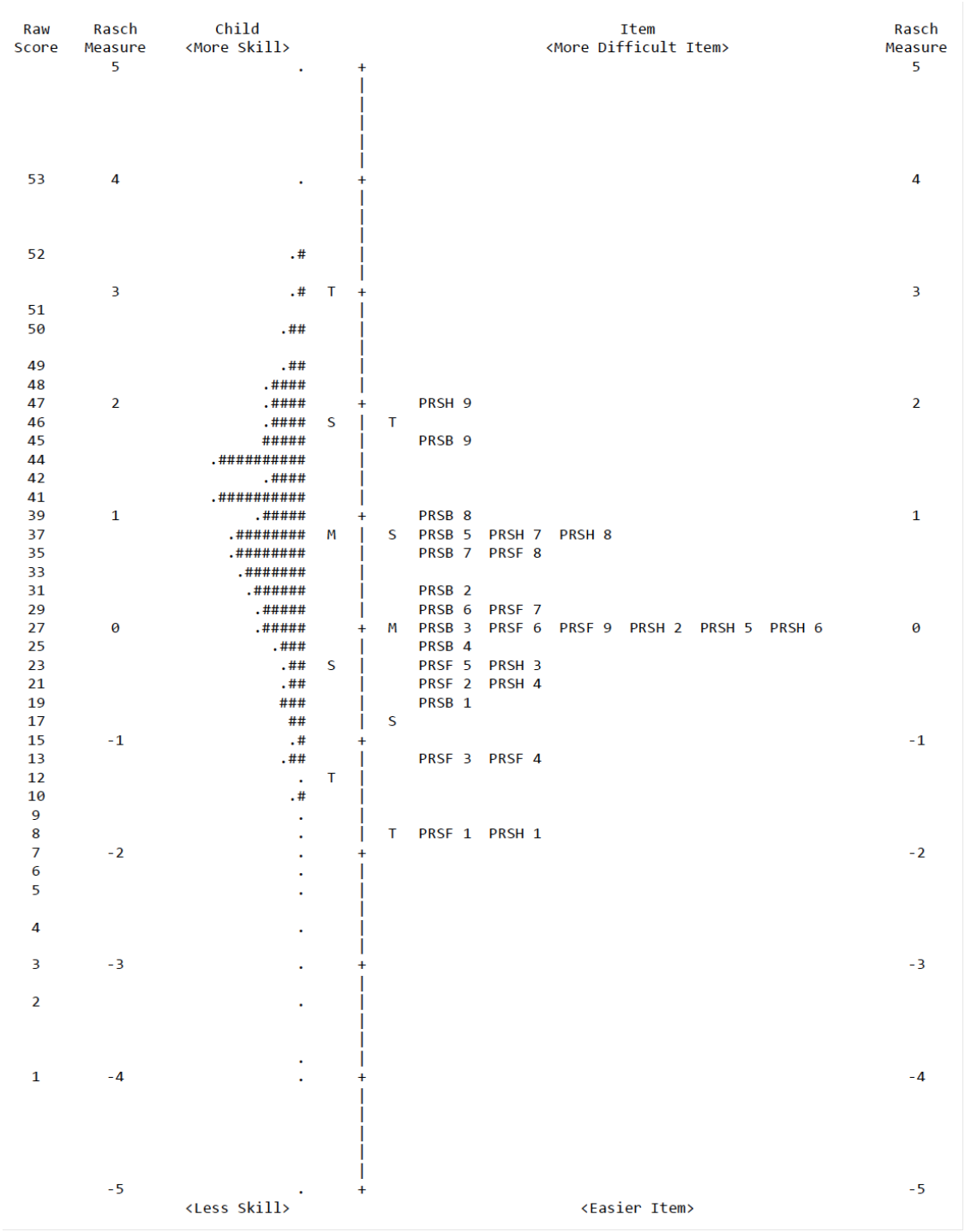



**Figure 3.2**

*Praxis: Sequences Wright Map*

*Note*. # = 20 children, . = 1 to 19 children, | = latent trait (praxis skill), M = item/child mean, S = 1 SD, T = 2 SD

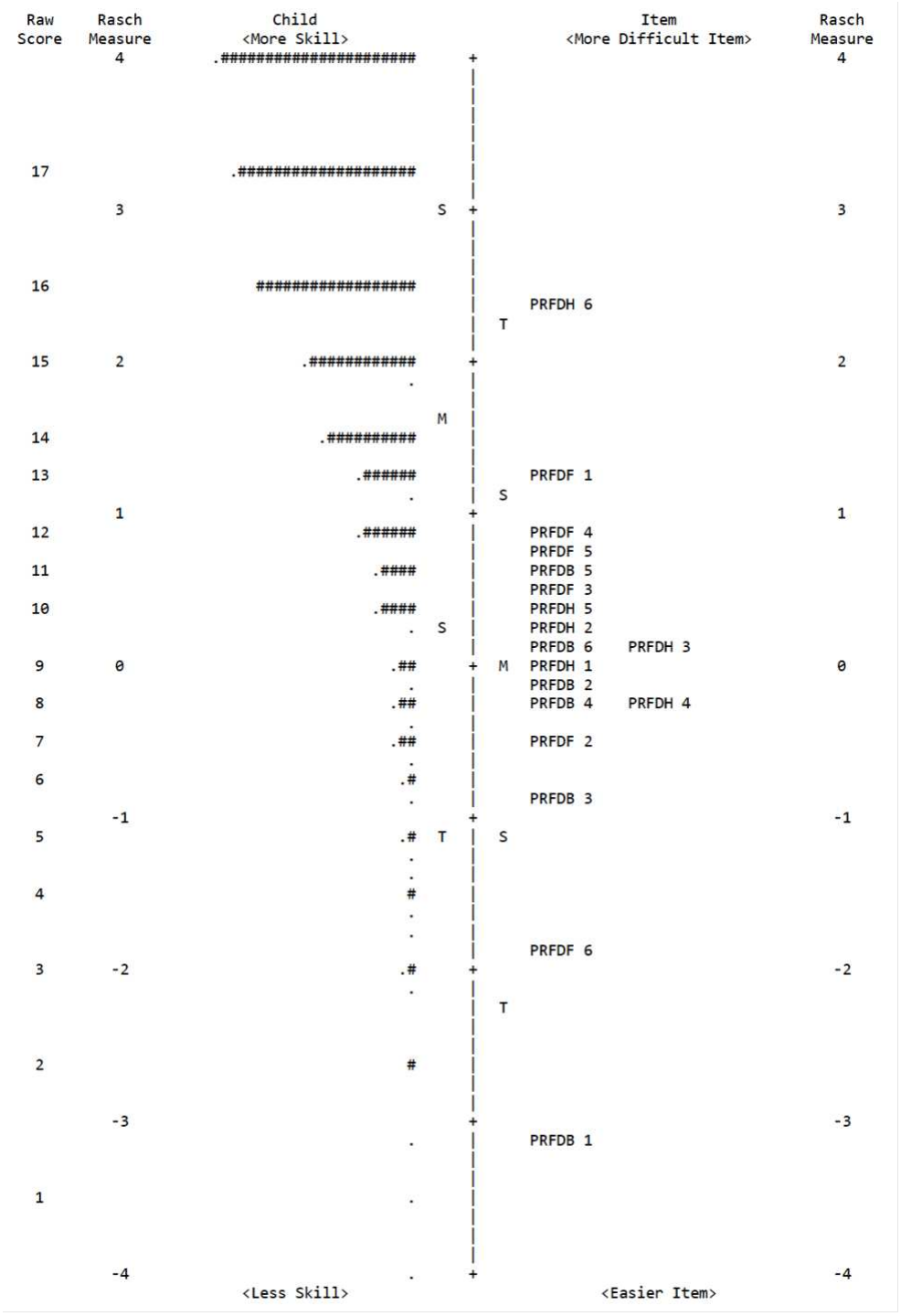



**Figure 3.3**

*Praxis: Following Directions Wright Map*

*Note*. # = 20 children,  = 1 to 19 children, | = latent trait (praxis skill), M = item/child mean, S = 1 SD, T = 2 SD

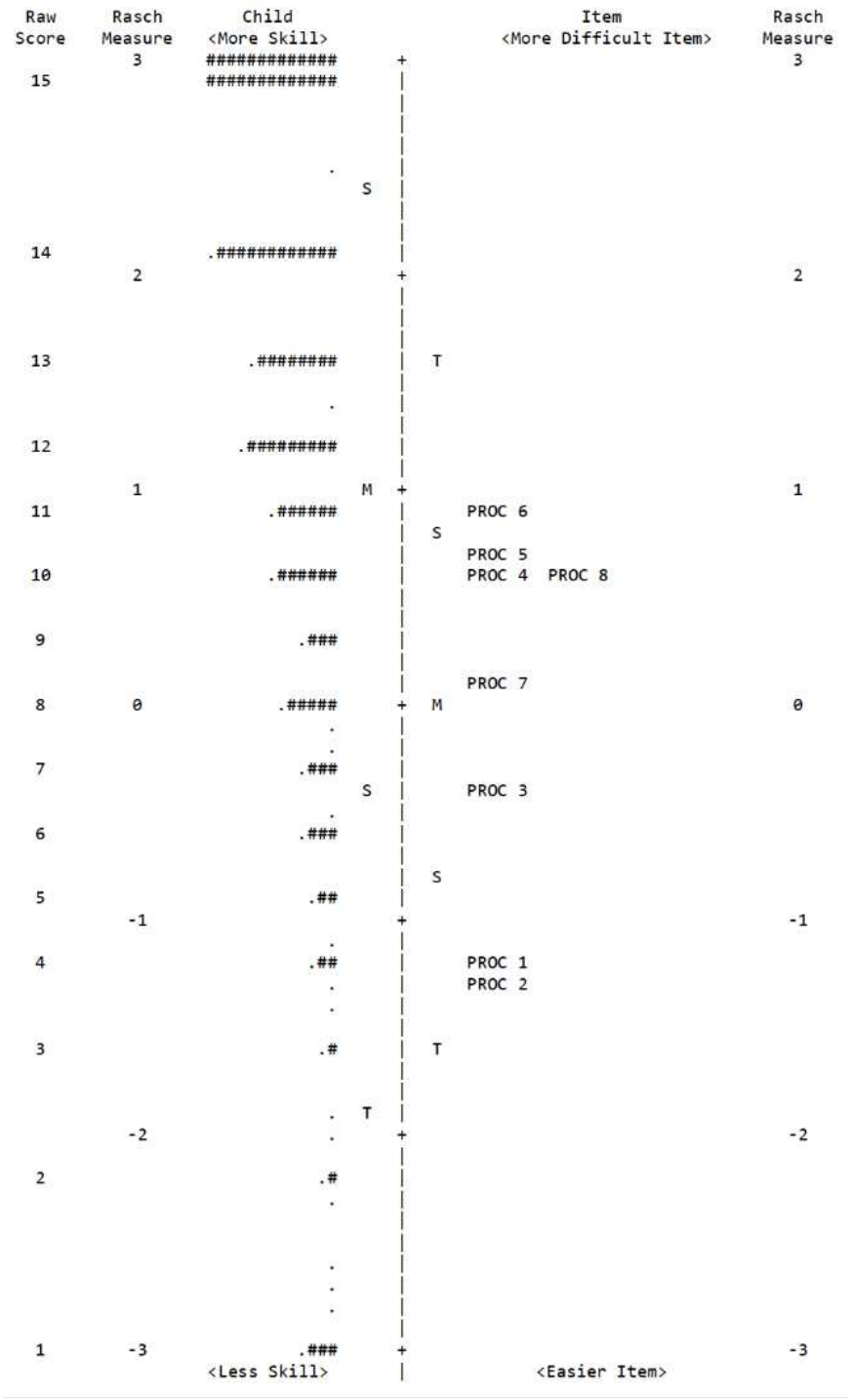



**Figure 3.4**

*Ocular Praxis Wright Map*

*Note*. # = 20 children, . = 1 to 19 children, | = latent trait (praxis skill), M = item/child mean, S = 1 SD, T = 2 SD

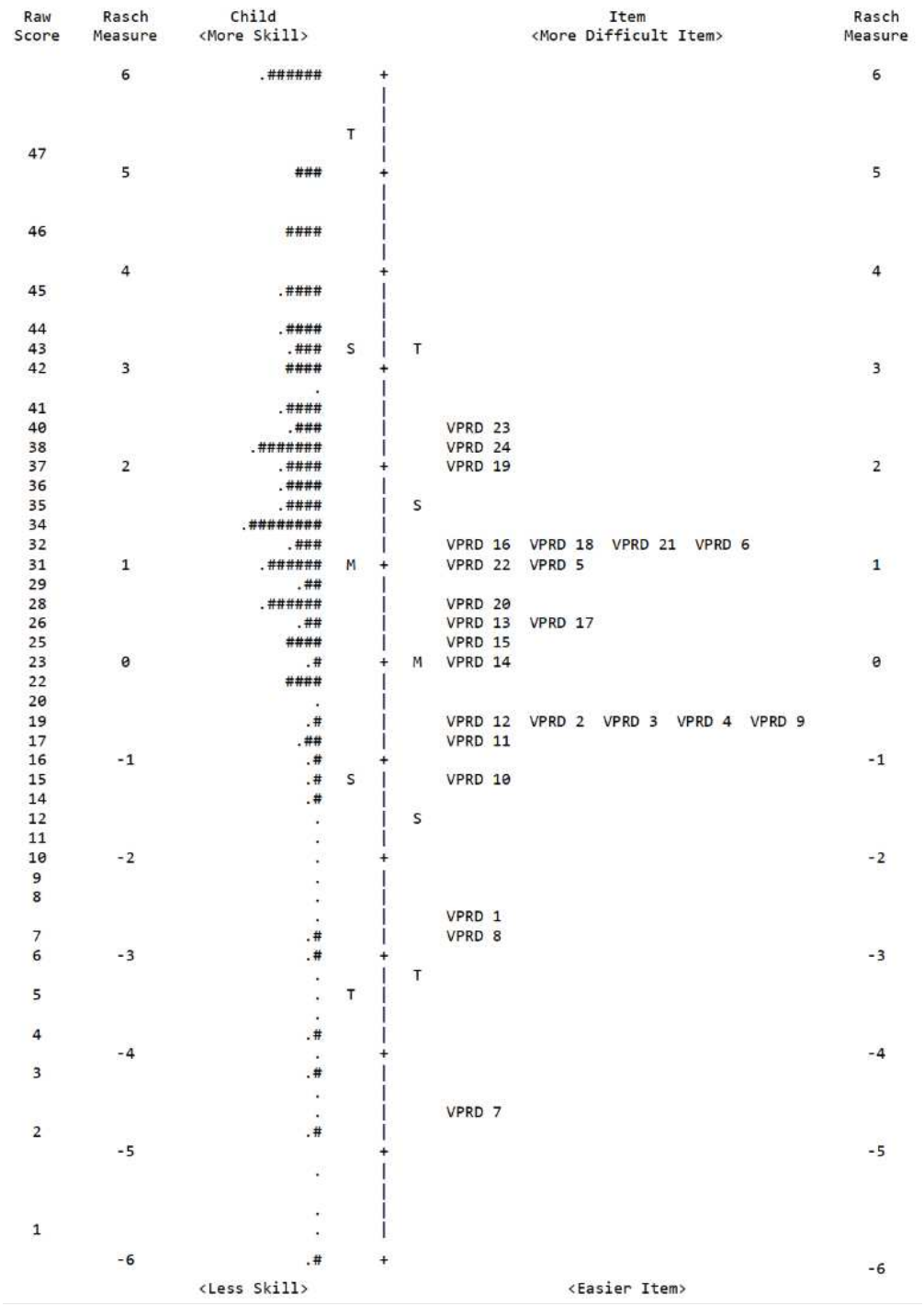



**Figure 3.5**

*Visual Praxis: Designs Wright Map*

*Note*. # = 20 children, . = 1 to 19 children, | = latent trait (praxis skill), M = item/child mean, S = 1 SD, T = 2 SD

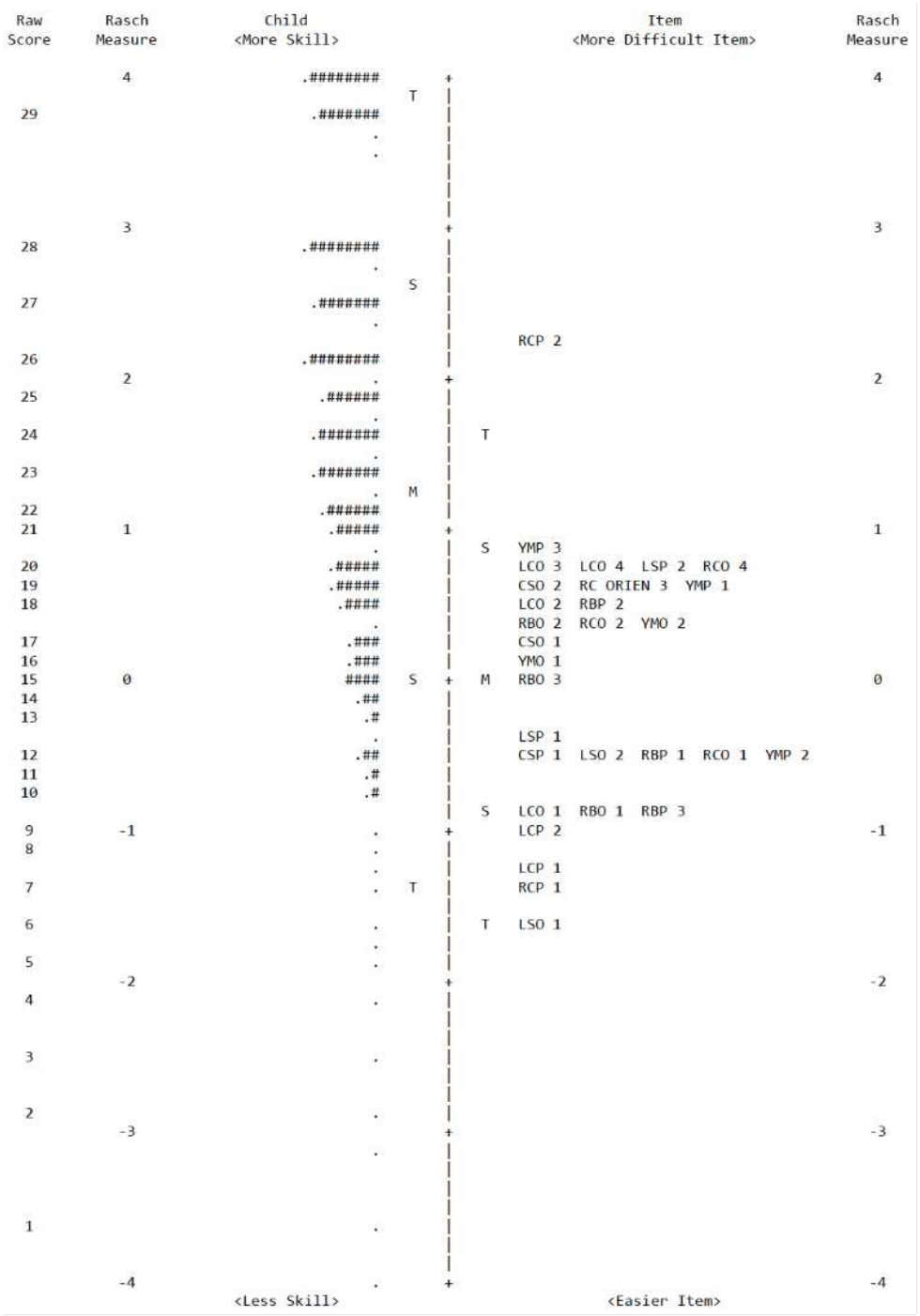



**Figure 3.6**

*Visual Praxis: Construction Wright Map*

*Note*. # = 20 children, . = 1 to 19 children, | = latent trait (praxis skill), M = item/child mean, S = 1 SD, T = 2 SD

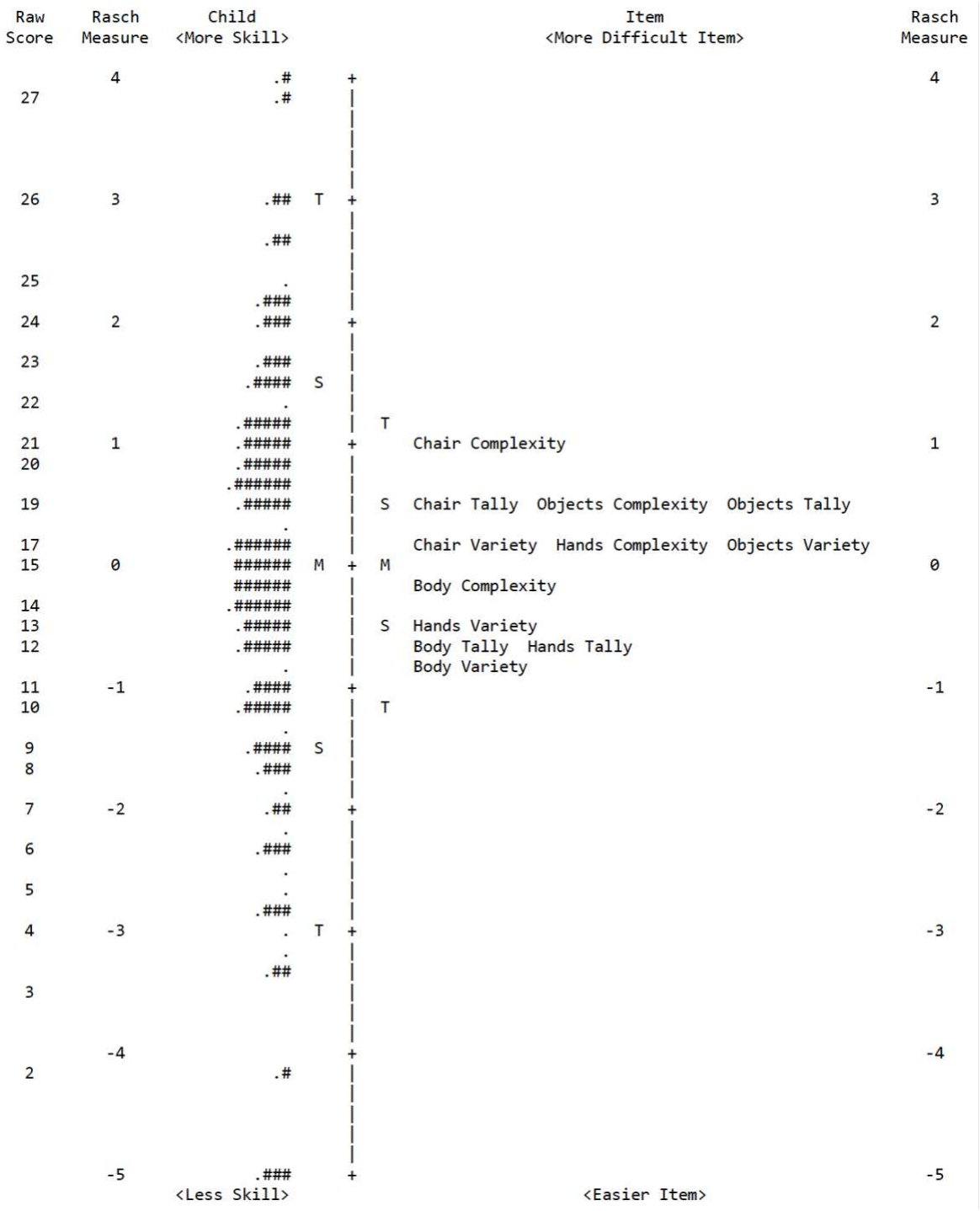



**Figure 3.7**

*Praxis: Ideation Wright Map*

*Note*. # = 20 children, . = 1 to 19 children, | = latent trait (praxis skill), M = item/child mean, S = 1 SD, T = 2 SD

**Table 3.12**

Mean Child Measure Scores

| Test | Child Mean (SE) |
|---|---|
| Praxis: Positions | 1.55 (0.46) |
| Praxis: Sequences | 0.84 (0.36) |
| Praxis: Following Directions | 2.14 (0.94) |
| Ocular Praxis | 1.72 (1.03) |
| Visual Praxis: Designs | 1.29 (0.57) |
| Visual Praxis: Construction | 1.48 (0.63) |
| Praxis: Ideation | -0.10 (-0.56) |

The person maps (Figures 3.8-3.14) demonstrated expected age progression in person measure at each level (i.e., older children tended to score higher on all praxis tests). Table 3.13 contains the results of Pearson correlations examining the relationship between age and Rasch-generated measure scores. All correlations were large ($\geq$.50; Cohen, 1988) except PrI, which was moderate ($\geq$.30).

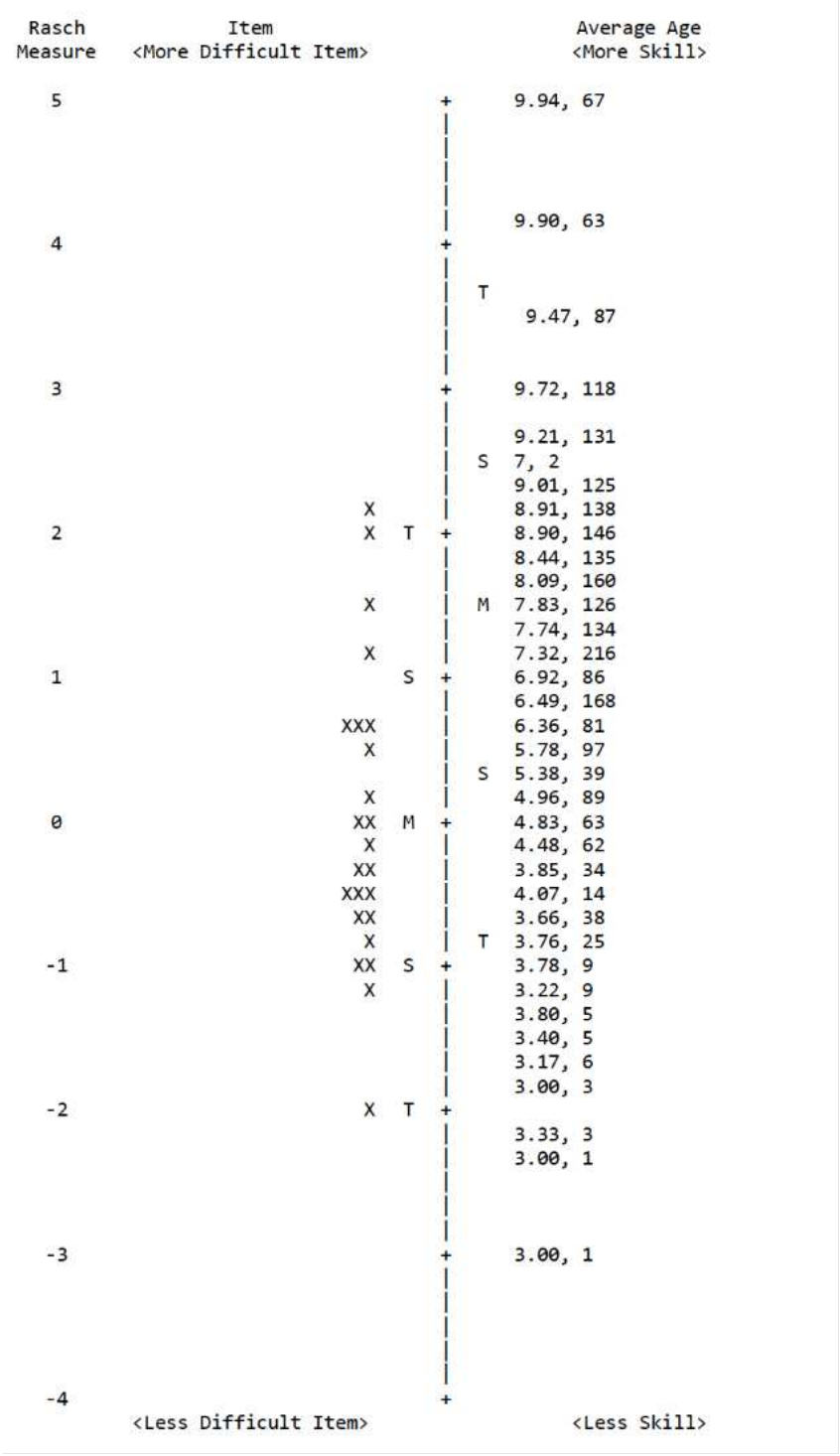



**Figure 3.8**

*Praxis: Positions Person Map*

*Note.* | = latent trait (praxis skill), left side of the figure shows items (X), right side of the figure shows (average age, # of children), M = item/child mean, S = 1 SD, T = 2 SD

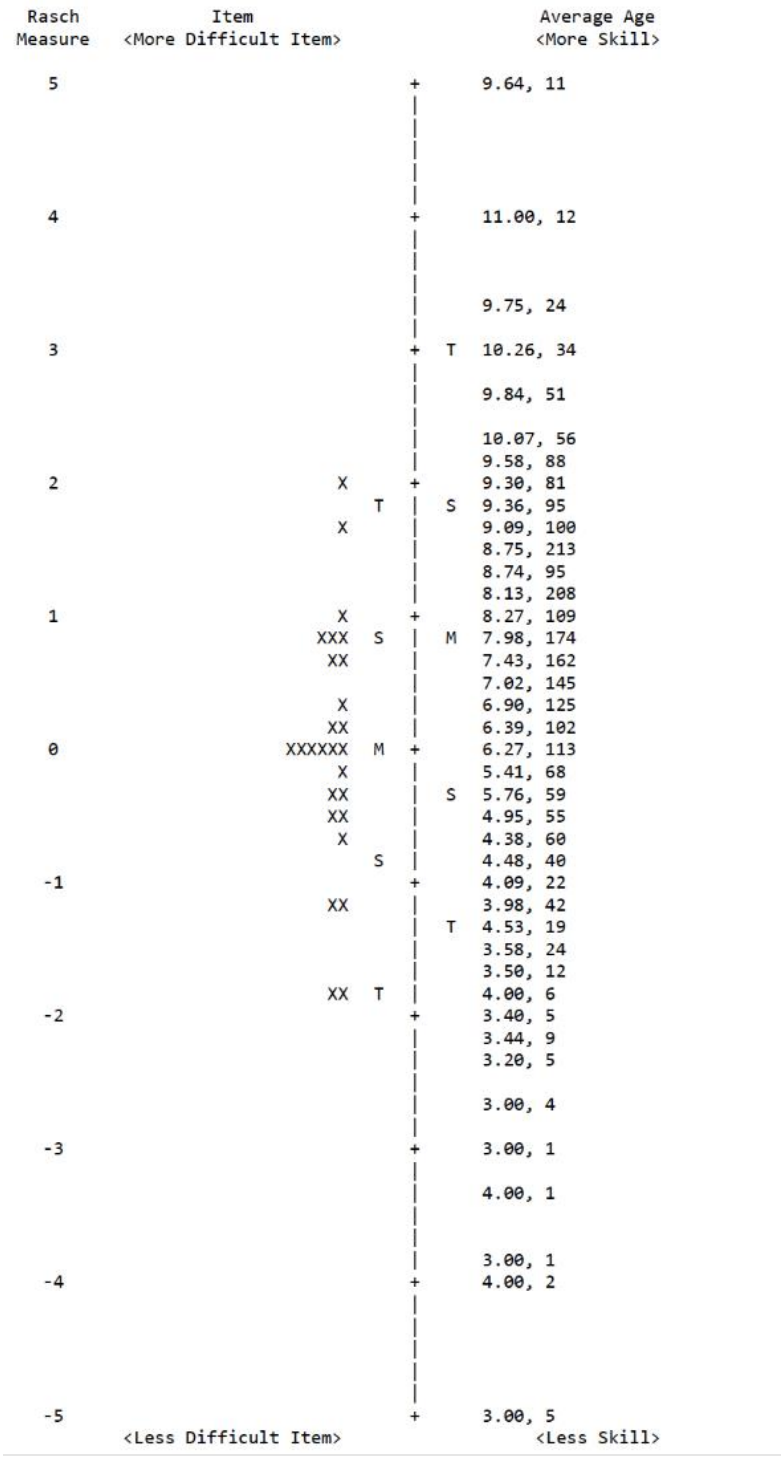



**Figure 3.9**

*Praxis: Sequences Person Map*

*Note.* | = latent trait (praxis skill), left side of the figure shows items (X), right side of the figure shows (average age, # of children), M = item/child mean, S = 1 SD, T = 2 SD

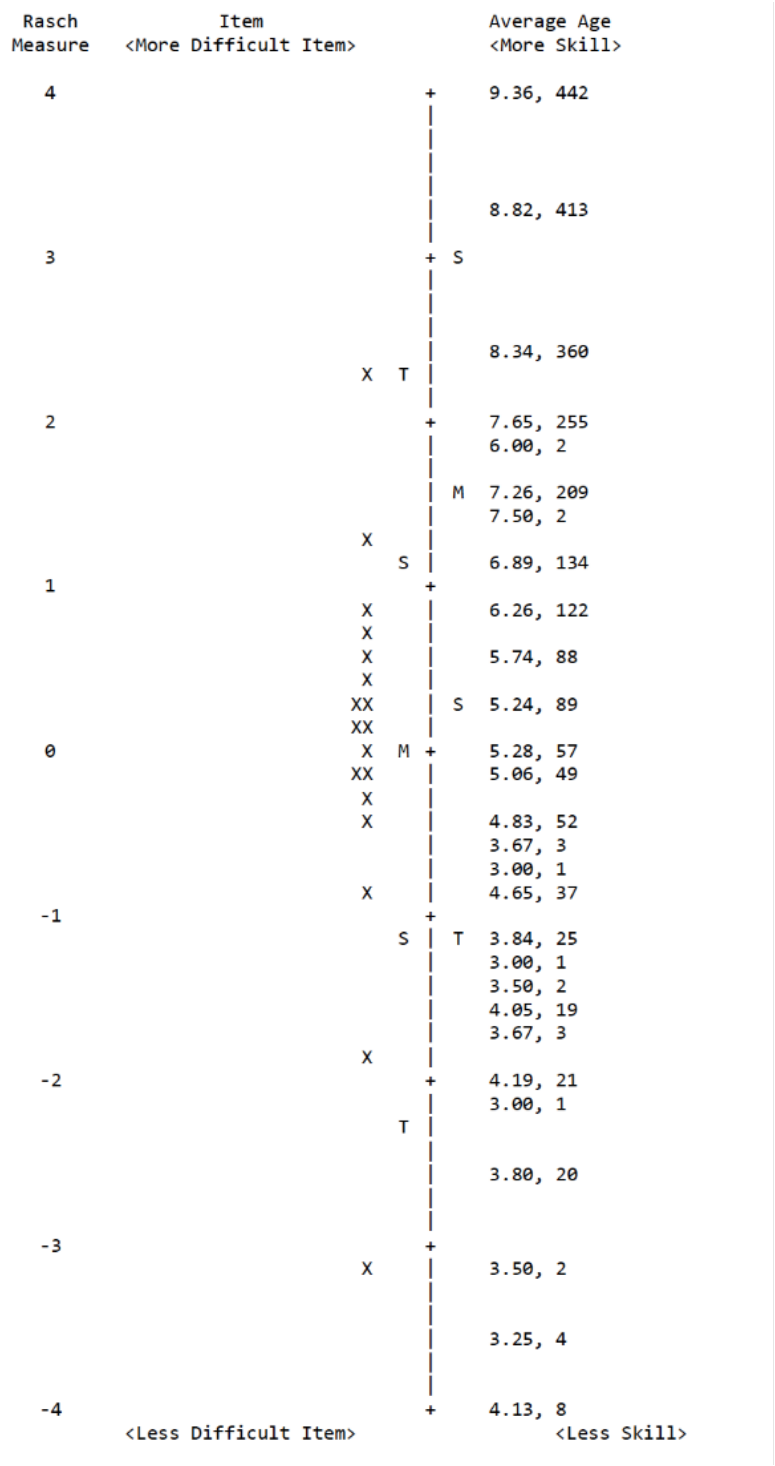



**Figure 3.10**

*Praxis: Following Directions Person Map*

*Note.* | = latent trait (praxis skill), left side of the figure shows items (X), right side of the figure shows (average age, # of children), M = item/child mean, S = 1 SD, T = 2 SD

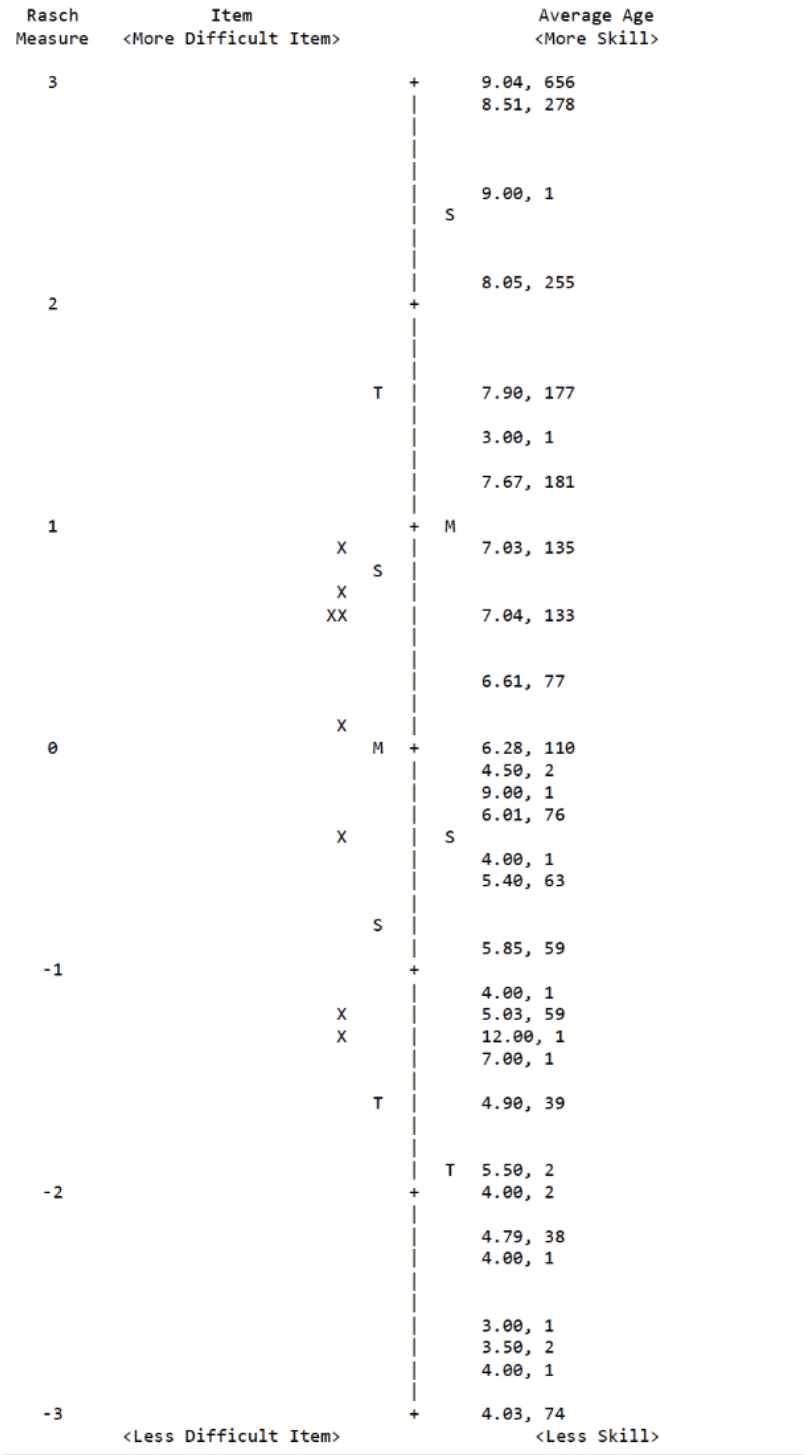



**Figure 3.11**

*Ocular Praxis Person Map*

*Note.* | = latent trait (praxis skill), left side of the figure shows items (X), right side of the figure shows (average age, # of children), M = item/child mean, S = 1 SD, T = 2 SD

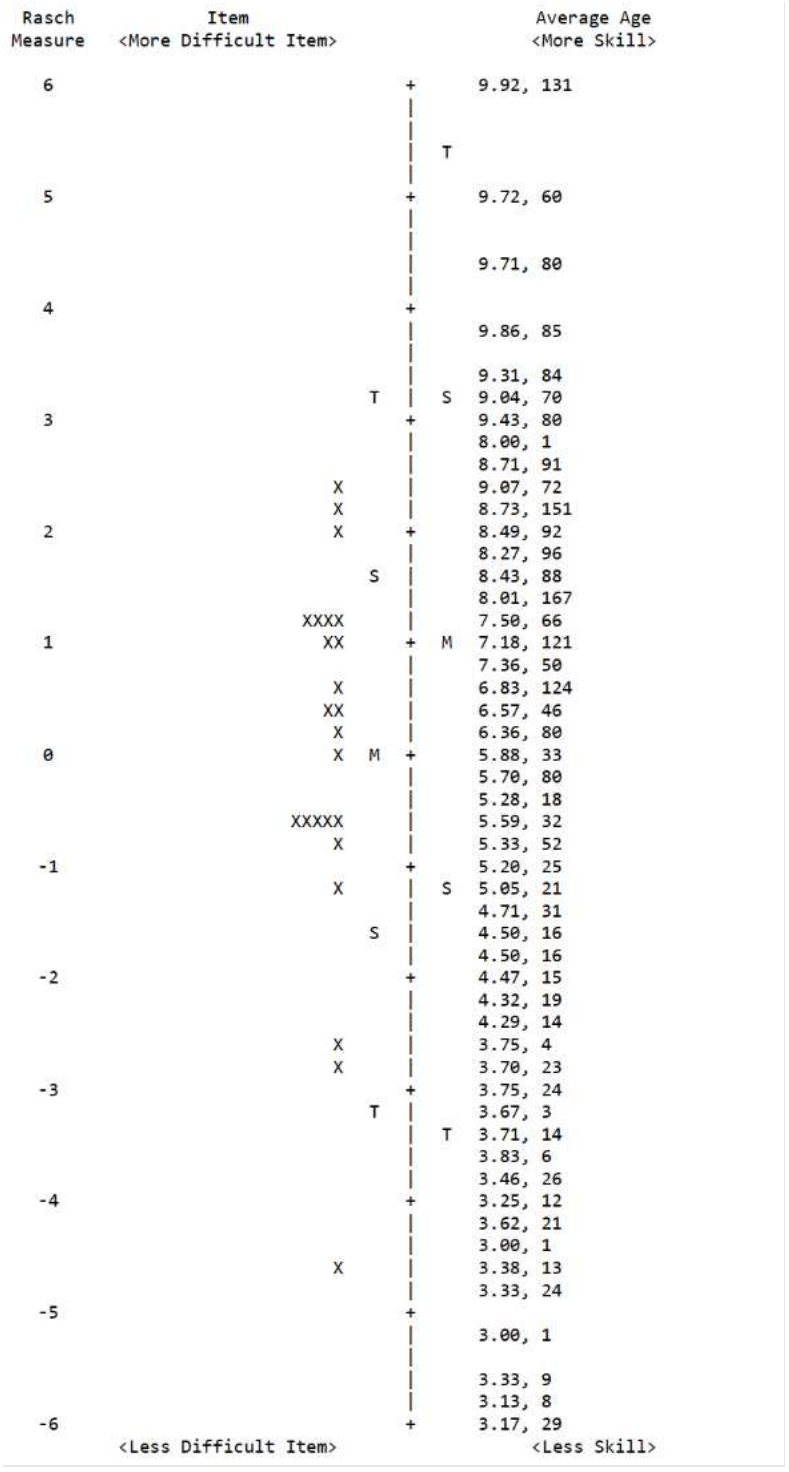



**Figure 3.12**

*Visual Praxis: Designs Person Map*

*Note.* | = latent trait (praxis skill), left side of the figure shows items (X), right side of the figure shows (average age, # of children), M = item/child mean, S = 1 SD, T = 2 SD

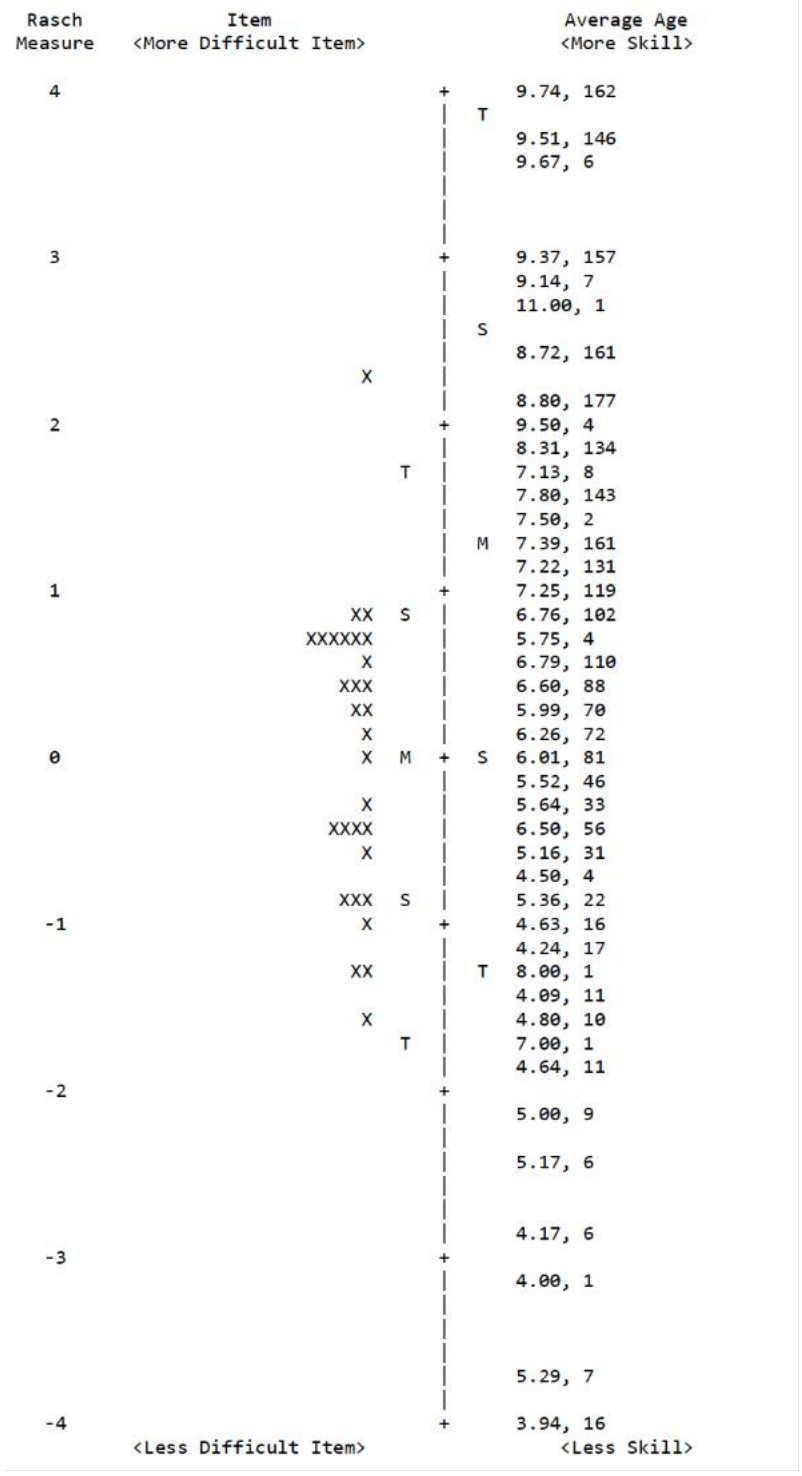



**Figure 3.13**

*Visual Praxis: Construction Person Map*

*Note.* | = latent trait (praxis skill), left side of the figure shows items (X), right side of the figure shows (average age, # of children), M = item/child mean, S = 1 SD, T = 2 SD

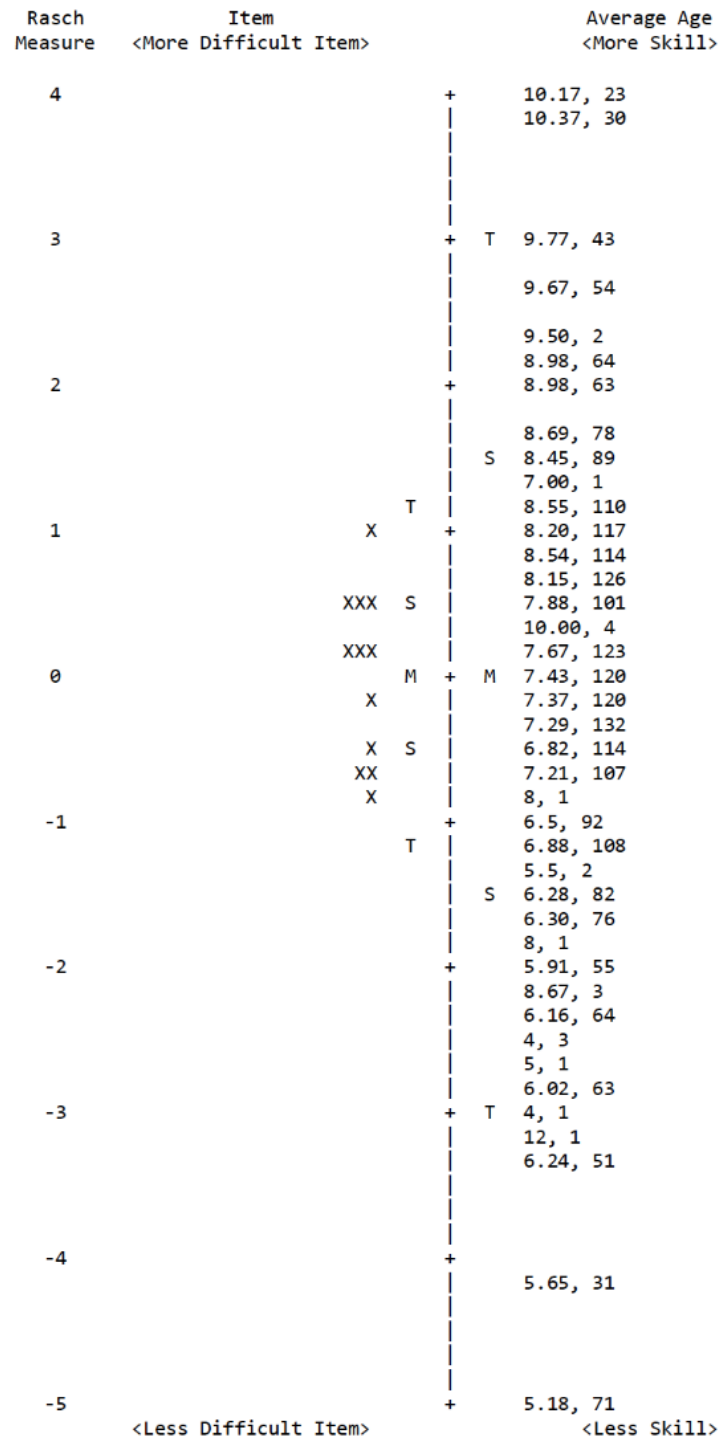



**Figure 3.14**

*Praxis: Ideation Person Map*

*Note.* | = latent trait (praxis skill), left side of the figure shows items (X), right side of the figure shows (average age, # of children), M = item/child mean, S = 1 SD, T = 2 SD

**Table 3.13**

*Correlation between Rasch Measure Score and Age in Months*

| Test | Correlation |
|---|---|
| Praxis: Positions | 0.64*** |
| Praxis: Sequences | 0.66*** |
| Praxis: Following Directions | 0.61*** |
| Ocular Praxis | 0.54*** |
| Visual Praxis: Designs | 0.74*** |
| Visual Praxis: Construction | 0.54*** |
| Praxis: Ideation | 0.43*** |

*Note.* *** = $p < .001$

**Internal Reliability**

Table 3.14 displays reliability indices (person reliability index and strata) for each of the seven praxis tests. All tests demonstrated adequate or better evidence for reliability based on both person-reliability index and strata. Notably, PrOc and PrFD did not achieve strong evidence based on strata (>3.0), although they did meet my minimum criteria of 2.0.

**Table 3.14**

*Rasch Reliability*

| Test | Person Reliability Index | Strata |
|---|---|---|
| Praxis: Positions | .84 | 3.44 |
| Praxis: Sequences | .89 | 4.12 |
| Praxis: Following Directions | .70 | 2.37 |
| Ocular Praxis | .74 | 2.60 |
| Visual Praxis: Designs | .95 | 5.93 |
| Visual Praxis: Construction | .82 | 3.15 |
| Praxis: Ideation | .88 | 4.00 |

## Discussion

I used Rasch analysis to examine evidence for validity and reliability of data collected with the seven EASI praxis tests. Goodness-of-fit statistics for both items and children suggested good adherence to the assumptions of the Rasch model – in other words, the items that comprise

each test formed unidimensional constructs of the underlying latent traits (praxis skills). Rating scales were adequate for all tests. PCA and DIF analyses provided evidence that off-dimensional noise in the data does not degrade measurement of praxis function. Reliability indices for all tests confirmed that measures could be reproduced with a similar sample of children.

**Model Fit and Unidimensionality**

While most items fit the Rasch model assumptions across all tests, I chose to retain those that failed to fit for several reasons. First, items may have failed to fit due to their administration order (PRPF1, VPRD7). On PrP, for example, PRPF1 is the first item that requires children to model a position with the face; the previous items asked children to assume hand positions. The sudden change in task activities may have impacted fit. This would likely be the case regardless of item content; therefore, I retained these items.

Other misfitting items (VPRD1, VPrC RCP2) occupied extreme item measures on the constructs (very easy and very difficult, respectively). Extreme items are more prone to misfit because aberrant responses are more unexpected (Wright, 1991). In the case of VPrD Design 1, for example, nearly all children could accomplish this item; therefore, incorrect responses were very unexpected and showed an outsized effect on fit statistics. I chose to retain these items because failure to fit was not severe. I suggest continued investigation of these items during future studies.

In general, the dot grid items of VPrD showed poorer fit than the other test items; four of the six items failed to fit while only one of 17 free-hand items did not fit the model. The dot grid items may measure a different aspect of visual praxis than the free-hand items; further – the dot grid provides a visually stimulating background that may negatively affect some children. The poor fit statistics may suggest that the items are unexpectedly easy for some children and

unexpectedly difficult for others. The visually stimulating background may have introduced more difficulty for some children, while others successfully used it as a guide. Therefore, these items may be best scored separately or omitted from total scores. At this time, I do not recommend removing these items; however, future researchers should closely monitor the dot grid items for evidence of misfit in clinical populations.

PCAs provided strong evidence for unidimensionality. No tests exceeded our criteria (both eigenvalue > 2.0 and disattenuated correlation < .57). However, it is worth noting that PrI had an eigenvalue of 2.94 and disattenuated correlation between item clusters 1 and 3 of .60. Further inspection of the items that comprise each cluster suggested that tally scores clustered together while quality scores (complexity and variety) formed another cluster. PrI is the first test measuring these quality constructs. The only other instrument designed to measure ideational praxis (TIP, Lane et al., 2014) scored ideation based solely on tally of novel ideas. Lane et al. did observe qualitative differences between older and younger children, especially in variety and complexity of actions. Cermak and May-Benson (2020) noted that children with dyspraxia driven by poor ideation may engage in repetitive or simple play. Therefore, I hypothesize that these items are a useful part of the construct of ideation. I suggest continued monitoring the PCA for these items during future studies, especially those including children with known dyspraxia.

Person fit analyses for five of the seven tests fell slightly below the desired criteria (fewer than 90% of children fit the model). Further inspection of children whose data failed to conform to the expectations of the model suggested that these children demonstrated poor outfit statistics but acceptable infit. In other words, these children had unexpected responses to items that were far from their ability level; children with poor fit largely showed unexpected incorrect responses to easy items. These responses may have been the result of inattention or boredom; shortening

the EASI tests may improve fit statistics. Additionally, I suspect that children with sensory integration concerns will have better fit statistics, as the items will be better targeted for these children's abilities.

The results of DIF analyses revealed no notable differences in item difficulty between male and female children. This is consistent with previous studies using the SIPT; these studies revealed no observable differences in praxis performance between male and female children (Ayres, 2005; Van Jaarsveld, Mailloux & Herzberg, 2012). Our study is the first to examine sex differences based on individual item difficulty.

Overall, fit analyses suggest strong evidence for construct validity of the seven EASI praxis tests; each test measures a unidimensional aspect of praxis (Bond, Yan & Heene, 2020). These findings agree with the pilot study findings for PrP, PrS, PrFD and PrI (Mailloux et al., under review). This is the first study to evaluate construct validity of data gathered with VPrD, VPrC, and PrOc.

**Item Distribution and Hierarchies**

Quantitative Rasch results, as described above, provide evidence for construct validity. However, fit statistics must be considered in context of the item hierarchies: the items that comprise a valid test should progress from easier (i.e., requiring *less* of the construct) to more difficult (i.e., requiring *more* of the construct) (Bond, Yan & Heene, 2020). To rigorously evaluate item hierarchies, the Rasch-generated construct should be compared with existing literature detailing the construct to be measured. Unfortunately, there is a dearth of literature evaluating praxis at this level of detail, preventing this analysis. Upon inspection, though, the Wright maps presented here (Figures 3.1-3.8) present logical hierarchies (described in Results).

Encouragingly, most tests contained items simple enough to capture the abilities of children who were more than 1.5-2.0 SD below the mean ability level (i.e., items with difficulty levels close to the children's ability levels). This suggests that these tests will be sensitive to differences even among the youngest children. However, some tests appear too easy to accurately measure children with the most developed praxis skill. PrOc is the most extreme example; hundreds of children receive logit scores as high as 3.0 while the most difficult item has an estimate measure of .92 logits. As a result, these tests may be less sensitive to differences among the most able children (Lunz, 2010). This discrepancy may be related to the population sampled for this study: all the children in this study were typically developing. Because the goal of this instrument is to identify and characterize the skills of children with sensory integration deficits, I am less concerned about poor sensitivity to children at the top of the scale. Future studies will determine the tests' sensitivities for detecting deficits in older children with clinical diagnoses.

Lending further support to construct validity, I found evidence that scores followed developmental trends, as expected. Across all tests, I observed moderate to strong correlations between age and Rasch-generated measure scores. Person maps also demonstrated progression in scores from the youngest to the oldest children. This is consistent with previous literature examining the SIPT, which suggested that praxis scores on imitation and verbal command improve with age (Ayres, 2005). The SIPT visual praxis tests (Design Copying and Constructional Praxis) also showed similar developmental trends to those observed in this study. Interestingly, our results for Praxis: Ideation diverge from the findings of Lane et al. (2014), who used TIP to evaluate children's abilities to generate novel ideas for interacting with objects. These authors found no differences in the number of novel ideas across 3-, 4-, and 5-year-olds.

Our results suggest a moderate correlation with age; this may be the result of our larger age range (3-12 years) or the more varied demands of the PrI test. While the TIP asks children only to generate ideas to play with a string, PrI asks children to show us what they could do with their bodies, hands, a chair, and a group of small objects. TIP only evaluates the quantity of ideas generated, while PrI scores children on the quantity, variety and complexity of their ideas. Lane and colleagues (2014) did observe more complex ideas from older children; therefore, PrI is likely more sensitive than TIP to developmental changes in ideational praxis.

Rasch reliability coefficients provided additional evidence that the tests targeted the sample population. High coefficients suggest that the spread of item difficulties matches the ability of the tested population (Linacre, 2022). It is encouraging that these coefficients met or exceeded our threshold for acceptable reliability for all tests. However, further studies of reliability should be conducted (e.g., test-retest reliability, inter-rater reliability) to better understand how these tests will perform in clinical contexts.

**Limitations**

Although this study included a relatively large, international sample, examiners recruited children based on convenience; many children were known to the examiners in advance of testing. This may have impacted results. Also, most examiners were clinicians working with children who have SI disorders. To prevent a skewed sample, our team asked examiners not to recruit siblings of children in their practice, as siblings may be more likely to share diagnoses and may not represent the typically developing population. However, this may have 'over-corrected' the sample and resulted in a sample with higher ability than the true typically developing population.

Our normative data collection efforts took place during the COVID-19 pandemic. As a result of varied restrictions across countries, our sample is not evenly distributed across the included world regions. Data collection is still underway. As the restrictions change, I expect more countries to meet their normative data collection goals, resulting in a more representative population. When sample sizes permit, DIF analyses should be completed by countries and/or world regions to examine whether norms should be stratified by location.

**Implications and Recommendations**

While these findings generally support construct validity and reliability of data gathered with the EASI praxis tests, item difficulty statistics suggest that many of the items are too easy for children with high-average praxis abilities (i.e., mean person abilities were substantially higher than item difficulties). Ordinarily, I might recommend additional items for these tests to ensure that older children with SI deficits are identified. However, because the EASI praxis tests are already quite long, I hesitate to recommend more items without removing others.

In the future, computer adapted testing (CAT) may allow examiners to administer only items that match the child's ability. CAT uses computerized algorithms to select only the items closest to the child's ability level, re-estimating ability each time the child responds to an item (Linacre, 2000; Cella et al., 2007). Using CAT, test developers can add harder praxis items to the item pool without increasing burden on test administrators. Unfortunately, CAT will not be appropriate for all tests; for example, for VPrC, the room design activity must be completed before evaluating individual items.

Alternatively (or in addition), the individual praxis tests may be combined to form longer, composite praxis evaluations. Lai et al. (1996) used Rasch analysis to demonstrate that four of the SIPT praxis tests (Sequencing Praxis, Oral Praxis, Graphesthesia and Postural Praxis), along

with Bilateral Motor Coordination, formed a unidimensional construct of praxis. Four of these tests are roughly equivalent to EASI tests (PrS, VPrD, PrP and BI [not examined in this study]). I can reasonably expect that the EASI praxis tests may also form a unidimensional construct. If future researchers find this to be true, each test may be shortened to a more parsimonious item set that captures more levels of difficulty without sacrificing clinical utility.

**Conclusions**

The findings presented here suggest that the EASI normative data are valid and reliable. The data form a robust basis for the EASI; based on normative scores, users can compare the results of children with suspected sensory integration deficits to children who are meeting developmental expectations. However, users should exercise caution when testing older children with suspected deficits, as some tests may be too easy to capture their praxis dysfunction. As always, clinicians should use standardized assessment data in conjunction with clinical observations, interviews and other instruments to determine the best intervention strategies for children with SI dysfunction.

# CHAPTER 4: MANUSCRIPT 2: SENSORY PERCEPTION

The ability to perceive and interpret sensations from the body and the environment is fundamental to participation in everyday occupations. In sensory integration (SI) theory, sensory perception forms the basis of motor and praxis skills; in turn, these skills allow children to engage with others and their environments through purposeful movement, exploration, and communication (Ayres, 1972; Bundy & Lane, 2020).

Sensory perception comprises layers of complexity. Lane and Reynolds (2020) defind several aspects. Sensory *detection* refers to the ability to perceive the existence of a sensation. This occurs at the peripheral level: at sensory receptors in the skin, eyes, and other sensory organs. Sensory *recognition* refers to the ability to identify the qualities of a sensation (e.g., the direction or intensity of touch). Recognition relies on both peripheral receptors and information processing in the central nervous system (CNS). Sensory *integration* refers to the ability to perceive and interpret sensations, to integrate multiple sensations into a meaningful whole, and to use sensory information to inform behavior; this occurs in the CNS. Although sensory perception spans several anatomical levels, in this paper, I use the term *sensory perception* to refer to the child's functional capacity to perceive the existence of stimuli and its qualities (detection and recognition). Other literature (e.g., Miller et al., 2007) uses the term sensory *discrimination*; for the purposes of this paper, I consider these terms interchangeable.

In sensory integration theory, sensory perception is thought to inform behavior by creating the child's internal sense of body shape and position as well as generating knowledge about the external environment. Children integrate somatosensory (tactile and proprioceptive), movement-related (proprioceptive and vestibular), and visual systems to create a "map" of their

bodies within their environments (Lane & Reynolds, 2020). This allows the child to predict and plan movements and actions (this ability is called *praxis*). Disordered praxis (dyspraxia), then, is thought to stem from impaired sensory perception – especially in the tactile, proprioceptive and vestibular (somatosensory) domains. The roles of other sensory systems (i.e., auditory, gustatory, and olfactory) in creating the body map is less clear, although emerging evidence suggests that humans may use auditory spatial mapping to inform body scheme as well (King et al., 2011).

Children with disordered sensory perception and dyspraxia may fail to develop age-appropriate motor planning and motor skills because the CNS does not receive accurate signals from the body and environment. Based on these neurological underpinnings, SI theory details an intervention approach for occupational therapists. SI intervention takes advantage of neuroplasticity in the CNS by providing opportunities for enhanced sensations that elicit adaptive responses (i.e., skilled responses to environmental challenges; Bundy & Szklut, 2020). However, SI intervention relies upon accurate assessment of sensory perception.

**Assessment of Sensory Perception**

To deliver high-quality SI treatment, therapists must begin by assessing the child's unique sensory perception challenges and abilities. Several standardized approaches may be used to quantitatively evaluate tactile perception, including monofilaments, aesthesiometers, and vibrometers (Hilz et al., 1998). However, few studies have examined the validity of these tools for pediatric populations. Moreover, they require specialized equipment and may not be feasible in pediatric clinical settings. There are no commonly used standardized tests of other somatosensory domains (i.e., vestibular and proprioceptive perception).

Assessment of visual perception relies upon standardized assessments such as the Motor Free Visual Perception Test (Colarusso & Hammill, 2015), the Test of Visual Perceptual Skills

(Martin, 2017), and the Developmental Test of Visual Perception (Hammill, Pearson & Voress, 2014). These tests are often used in pediatric clinical settings, although normative data are limited to the US population. Tests of auditory perception usually require specialized instrumentation and are not often delivered in occupational therapy clinics.

The Sensory Integration and Praxis Tests (SIPT; Ayres, 2005) provide comprehensive evaluation of sensory perception through seven tests. Five of these tests evaluate somatosensory perception: Manual Form Perception (stereognosis), Finger Identification (two-point discrimination, recognition and localization of static tactile stimuli), Localization of Tactile Stimuli (ability to identify where on the body one is touched), Graphesthesia (replication of dynamic tactile stimuli), and Kinesthesia (conscious proprioception). Two tests examine visual perception: Figure Ground (detection of a visual stimuli on a crowded background) and Space Visualization (matching a shape to its negative form). Each test provides a z-score; scores that are less than -1.0 indicate dysfunction. However, SIPT have several notable limitations. First, they were developed and normed in the 1960s and 1970s with a US-based population; therefore, normative data are limited in scope. SIPT also do not evaluate auditory perception. Finally, the required materials for each test are carefully standardized and can only be produced by the test manufacturer; as a result, the SIPT is prohibitively expensive for many underfunded clinics.

**The Evaluation in Ayres Sensory Integration**

The Evaluation in Ayres Sensory Integration (EASI; Mailloux et al., 2018) is a novel, norm-referenced, performance-based suite of instruments used to assess praxis, sensory perception and sensory reactivity in children ages 3-12 years. EASI are fully integrated with the constructs of SI theory; assessment results can be translated to suggest intervention approaches and methods. The EASI comprise 20 individual tests; eight of these tests evaluate sensory

perception in the tactile, proprioceptive, vestibular, visual and auditory domains. Table 4.1 describes the EASI sensory perception tests.

**Table 4.1**

EASI Sensory Perception Tests

| Test | Description | Scoring | # of Items |
|---|---|---|---|
| Tactile Perception: Localization (TPL)[1] | Examiner touches a spot on the child's arm or hand with either one or two fingers; child identifies the spot where they were touched and the number of spots they felt | 1: Child correctly identifies location (within 1cm) and number of stimuli<br>0: Child does not correctly identify location and/or number of stimuli | 20 |
| Tactile Perception: Designs (TPD)[1] | Examiner traces a design on the child's arm or hand with finger, child copies the design | 2: Child correctly replicates design<br>1: Child replicates design with minimal errors<br>0: Child does not correctly replicate design | 24 |
| Tactile Perception: Shapes (TPS)[1] | Items 1.1-1.8: Child matches a shape placed in the hand with a visual model on a field of distractors<br><br>Items 2.1-2.12: Child matches a shape or textured tile placed in the hand with an identical shape on a field of distractors | 1: Child selects correct model<br>0: Child does not select correct model | 20 |
| Tactile Perception: Oral (TPO)[1] | Child matches a shape presented to the mouth with a visual model | 1: Child selects correct model<br>0: Child does not select correct model | 11 |
| Visual Perception: Search (VPS)[1] | Child locates a visual stimulus on one of 3 visually crowded forms within 10, 20 or 30 seconds (dependent on form) | 1: Child locates correct object within allotted time<br>0: Child does not locate correct object within allotted time | 18 |
| Auditory Localization (AL)[1] | Clicker Items: Examiner presses a clicker next to various parts of the child's body; child identifies where they heard the sound<br><br>Table Items: Examiner taps underneath a table one or two times; child identifies which quadrant of the table was tapped and how many taps they heard | 1: Child identifies correct location and correct number of stimuli<br>0: Child does not identify correct location and/or correct number of stimuli | 20 |

| Test | Description | Scoring | # of Items |
|---|---|---|---|
| Proprioception: Joint Position (PJP)[1] | One Hand Items: Examiner places child's finger on various spots; child returns the limb to the same spot<br><br>Foot Items: Examiner places child's toe on various spots; child returns the limb to the same spot<br><br>Two Hands Items: Child matches the position of two hands at the same time on either side of an open door | **One Hand and Foot Items**<br>2: Child places finger/toe 0-2cm from target<br>1: Child places finger/toe 3-5cm from target<br>0: Child places finger/toe 6 or more cm from target<br><br>**Two Hand Items**<br>2: Child's hands are positioned 0-1 cm from each other vertically on each side of the door<br>1: Child's hands are positioned 2-4 cm from each other vertically on each side of the door<br>0: Child's hands are positioned 5 or more cm from each other vertically on each side of the door | 15 |
| Proprioception: Force (PF) | Crayon Items: Examiner makes a mark with a crayon; child makes a mark that matches the intensity of the examiner's mark<br><br>Bottle Items: Child rolls a bottle of rice to the same segment along a marked yoga mat two times, using one hand, two hands, and one foot | **Crayon Items**<br>2: Child's image closely matches intensity of examiner's image<br>1: Child's image has slight difference in intensity<br>0: Child's image does not match intensity of examiner's image<br>**Bottle Items**<br>2: Second attempt lands within the same segment or 1 segment away from the first attempt<br>1: Second attempt lands 2-4 segments from the first attempt<br>0: Second attempt lands five or more segments from the first attempt | 10 |

[1]Vision occluded during each of these tests

In this report, I used the Rasch model to evaluate evidence for validity and reliability of normative data collected for eight of the sensory perception tests of the EASI. The Rasch model is a subset of item response theory (IRT) which allows examination of individual items and their relationship to the overarching latent traits measured by each test (Bond, Yan, & Heene, 2020). Rasch analysis provides evidence for construct validity (i.e., that the items measure the constructs they are designed to assess (Bond, Yan & Heene, 2020). Specifically, Rasch analysis evaluates unidimensionality – that the items measure a single, coherent latent trait (e.g., ideational praxis). Furthermore, the Rasch model provides evidence for internal reliability (i.e., the precision of person measures).

Several of the EASI sensory perception tests have been subjected to psychometric analyses using the Rasch model. In a previous study, the EASI development team (including myself) conducted a Rasch analysis of the pilot versions of the four tactile perception tests with a group of children from the US with and with not known SI disorders (Schaaf et al., in press). We found evidence that supported the unidimensionality of these tests; 95.8% of items across the four tests showed adequate fit to the model, and the items were well-distributed across the latent variable. Reliability coefficients were adequate (i.e., person reliability indices > .70 and strata > 2.0) for TPD, TLP and TPS. TPO showed adequate strata but person reliability index of .69, just below our desired threshold. Further, Rasch-generated child measures for all tests differed significantly between the typical children and those with known SI dysfunction (all *p*s < .001). PJP and PF have also been subjected to preliminary analyses (Mailloux et al., 2021). All items fit the Rasch model. PJP showed adequate reliability (person reliability index = .83), while PF did not (person reliability index = .51). Group comparisons revealed significant differences between

children with and without known SI disorders for all PJP items except two hand items ($p$ = .057) and for the crayon items of the PF test.

These analyses led to substantial revisions to the tactile tests (primarily removing redundant items). Although the final manuscripts reported the validity of the revised tests, data were collected using pilot versions of the EASI tests (Mailloux et al., 2018). These tests were significantly longer; test fatigue and distraction may have impacted findings. New analyses should be conducted using the revised tests. Furthermore, AL and VPS have not yet been subjected to rigorous analyses.

The present study uses a large, international normative sample to establish the psychometric properties of the revised EASI tests using the Rasch model. These data will form the basis for generating normative scores (e.g., $z$-scores) for use in therapy clinics. Therefore, it is critical to demonstrate validity and reliability of these data. Moreover, the results of these analyses may lead us to suggest additional revisions to the EASI tests before they are available globally. Specifically, I examined the following research questions:

(1) What is the evidence for construct validity of the data collected using each of the 8 EASI sensory perception tests?
    a. Do the test items demonstrate uniformly positive point-measure correlations (i.e., do scores on each item correlate with overall test score?)
    b. Do 95% of items demonstrate adequate fit to the Rasch model?
    c. Do 90% of children demonstrate adequate fit to the Rasch model?
    d. Do the Rasch-generated step thresholds within rating scales progress in an orderly fashion?

e. Does a Rasch principal components analysis of standardized residuals (PCA) reveal meaningful secondary dimensions in the data?

f. Does a differential item functioning (DIF) analysis reveal invariance in item difficulty for male and female children?

g. Do the items form a logical hierarchy with sufficient item difficulty variation to match sample ability levels?

h. Do test-takers form a logical developmental hierarchy (i.e., do scores increase with increasing age?)

(2) What is the evidence for internal reliability of data collected using the 7 EASI praxis tests?

a. Does the data demonstrate adequate internal reliability based on the Rasch person reliability index?

b. Does the data reliably distinguish at least two levels of sensory perception, based on the number of strata associated with the measure?

## Methods

I used Rasch analysis to evaluate evidence for validity and reliability of international normative data collected using the eight sensory perception tests on EASI. I conducted a separate analysis for each of the tests; this report summarizes the results of these analyses.

### Participants

I drew data from the EASI International Normative Data Collection Project (maintained by the Collaboration for Leadership in Ayres Sensory Integration [CLASI]). The dataset comprises 2563 children between the ages of 3-12 years. Inclusion criteria for normative data collection included: (1) chronological age between 3 years 0 months – 12 years 11 months; (2)

typical development; (3) no known medical, educational, mental health, or other developmental concerns. Exclusion criteria were: (1) known medical, educational, mental health, or other developmental concerns; (2) identified as having sensory integration concerns by OT, PT, or SLP; (3) receive(s/d) therapy services for learning disorders, ASD, ADHD, speech/language delays, regulatory issues, hypotonia, or DCD; and (4) siblings who met any of these exclusion criteria. Not all children completed every test; individual test sample sizes are reflected in Table 4.10. In Appendix C, I described the entire sample. The Rasch model is robust against missing data; therefore, I included children who did not complete all items within tests. I only omitted children with less than 50% of items completed.

**Procedure**

All normative data collectors (examiners) completed an 8- to 10-hour online training course that covered EASI testing and scoring. At the conclusion of training, they completed a series of online scoring quizzes, achieving at least 80% accuracy against scores completed by a gold standard observer. Examiners conducted EASI tests in locations convenient for children and families. This included clinics, children's homes, and research laboratories. Most examiners gave all 21 tests during a single 3- to 4-hour session; however, some required multiple sessions. Most common reasons for multiple sessions included scheduling conflicts or children's limited tolerance for extended testing. Examiners uploaded all data into a secure RedCap database managed by CLASI.

**Data Analysis**

I conducted all data analyses using Winsteps (Linacre, 2022), a Rasch-specific software program. The Rasch model is a latent trait psychometric model that converts ordinal-level data (e.g., EASI raw data) to interval-level measures (Bond, Yan & Heene, 2020). Both person ability

and item difficulty are estimated along the same log-odds unit ("logit") scale. Rasch is based upon two complementary assumptions; these assumptions can be expressed in terms of the EASI sensory perception tests: (1) that easier items (i.e., items that require less precise sensory perception) are easier for all children, and (2) that children with more well-developed sensory perception will successfully complete harder items compared with children with less well developed sensory perception.

**Model Selection**

The Rasch model includes several sub-models, including the dichotomous model (DM) and the rating scale model (RSM) (Bond, Yan & Heene, 2020). I employed DM to estimate item and person parameters for the tests with only dichotomous items (TPL, TPS, TPO, VPS, AL). For the tests with trichotomous rating scales (TPD and PJP), I used RSM. RSM specifies that all items within a test share a common rating scale, but may have polytomous rating scales (Wright, 1998). In addition to item difficulty and person ability estimates, RSM provides logit calibrations for rating scale categories. For PF, the first six items are scored using one trichotomous scale, while the next six items are scored using a different trichotomous scale. Therefore, I selected the grouped RSM in which items within a test share multiple common rating scales. Like RSM, this model provides fit statistics and logit calibrations for rating scale categories.

**Construct Validity**

In addition to generating item, person and rating scale calibrations, Winsteps provides several indicators that suggest the extent to which the data fit the Rasch model. I examined these indicators for evidence of construct validity: point-measure correlations, goodness-of-fit statistics, rating scale thresholds (for polytomous scales), item hierarchies, PCA of standardized residuals, and DIF statistics.

***Point-measure Correlations.*** To determine if each item corresponds with the latent variable (i.e., that a higher score corresponds to improved sensory integration abilities), I examined Pearson point-measure correlation coefficients between observations and item measure. In the Rasch model, positive point-measure correlations suggest that items align with the construct (Bond, Yan & Heene, 2020). Of note, the magnitude of these correlations is less important than the directionality. To establish construct validity, all point-measure correlations should be positive.

***Goodness-of-fit Statistics (Items).*** I examined two kinds of mean-square goodness-of-fit-statistics generated by Winsteps: infit and outfit. Infit statistics are "inlier-sensitive" or information-weighted to reduce the influence of off-target responses (i.e., people whose overall scores are far from the item measure). Outfit statistics are unweighted and typically reflect fit problems due to outliers. Mean-squares show the amount of distortion of the measurement system. Ideal mean-square value is 1.0. Values between 0.7 and 1.3 suggest adequate fit to the Rasch model (Linacre, 2002).

To demonstrate sufficient evidence for construct validity, at least 95% of items on each test should show adequate fit to the Rasch model. For tests with fewer than 20 items, a single misfitting item would fall below this threshold. Because I might expect at least one item to fail to fit due to chance alone, I expected *either* 95% of items to fit the model, or *no more than one* item to misfit.

***Goodness-of-fit Statistics (Children).*** Person fit statistics are calculated and interpreted in the same way as item fit statistics. As a rule, people often behave less predictably than items (Bond, Yan, & Heene, 2020). Further, given that I had relatively few items compared to people, I selected less stringent criteria for acceptable mean-squares (0.5 to 1.5). People who overfit the

model (i.e., behave too predictably, mean-square < 0.5) are unlikely to distort or degrade the measurement system (Linacre, 2015). Therefore, I only considered children's data to fail to fit if they underfit the model (MnSq > 1.5). I expected that 90% of children would fit the model for each test for evidence of strong construct validity.

***Rating Scale Analysis.*** For all rating scales, I examined (1) rating scale goodness-of-fit statistics. Mean-square between 1.3 and 0.7 suggested fit to the Rasch model; Linacre, 2021). (2) observed average person measure associated with each category. The observed average person measure associated with each category should demonstrate orderly progression; the lowest category should correspond with the lowest average person measure (Bond, Yan & Heene, 2020). Finally, for tests with polytomous rating scales, I examined (3) Andrich thresholds (i.e., the person ability measure at which a person is equally likely to use two adjacent categories). Andrich thresholds should progress in an orderly fashion, such that the lowest step threshold corresponds to the threshold between the two lowest categories and so forth (Bond, Yan & Heene, 2020; Linacre, 2018). Thresholds are not calculated for dichotomous scales.

For PJP and PF Part 2 (Rolling Bottle) items, I created rating scales from difference items. For both tests, I computed differences (difference between the child's point and the target, for PJP; difference between the first and second bottle roll, for PF). Then, I reverse-scored so that a lower difference represented a higher score. Based on inspection of the data and iterative Rasch analyses, I collapsed the rating scales into those represented in Table 4.11.

***Item Hierarchy.*** I assessed the item hierarchies in two ways. First, I compared the mean item measure and the mean person measure. In the Rasch model, the mean item measure is set at 0.0 logits. Mean person measure close to 0.0 indicates a match between the sample's sensory integration ability and the scale difficulty (Bond, Yan & Heene, 2020). Second, I visually

inspected the Winsteps-generated Wright maps. The Wright map provides a hierarchy of items and persons along a logit scale, ranging from lowest to highest measures. These items should be ordered logically so that theoretically-more-difficult items are associated with higher item difficulty measures. The most robust way to examine the logic of an item hierarchy is to compare the items with existing literature; however, few previous studies have examined sensory perception items in the level of detail presented by the EASI. Therefore, I descriptively examined the item hierarchies to ensure that they matched theoretical expectations. I also examined the Wright maps to evaluate the spread of items; large gaps in item difficulty indicate a need for more items, while items grouped together suggest redundancy.

***Person Hierarchy.*** I assessed the person hierarchies in two ways. The Winsteps-generated person maps show child scores along a continuous, interval-level scale on the right side of the figure and items along the left side. The interval scale is broken into .2 logit levels. I averaged the ages of children on each level and visually inspected the map for evidence that average age increases with increasing scores. Second, I evaluated the strength of this relationship by conducting bivariate Pearson correlations between Rasch-generated child measure scores and children's age in months. Given the developmental nature of sensory integration constructs, I expected at least moderate correlation coefficients ($\geq$.30; Cohen [1988]). I confirmed normality of all variables (age in months and EASI measure scores) using the methods described by Kim (2013) for large sample sizes ($N > 300$).

***Principal Components Analysis.*** While goodness-of-fit statistics and other evidence described above reflect the extent to which a construct is unidimensional, PCA provides evidence of the strength of additional dimensions in the data (i.e., multidimensionality). PCA deconstructs model residuals to identify additional dimensions in the data (i.e., item response

patterns not explained by the Rasch model). Eigenvalues estimate the strength of these dimensions (called contrasts). I considered contrasts to be strong enough to refute unidimensionality of the construct if the following conditions are met (Linacre, 2018): (1) There are contrasts with eigenvalues > 2 (i.e., with the strength of more than 2 items); (2) Item subsets within contrasts demonstrate disattenuated correlations < 0.57, indicating that item subsets likely measure different latent variables. According to Wright and Stone (1979), unidimensionality is essential to good construct measurement; evidence of multiple dimensions suggests that the items should be scored as multiple, separate instruments.

***Measurement Invariance.*** I used Rasch differential item functioning (DIF) analyses to examine the measurement invariance of the perception tests based on sex (i.e., that test items are not biased based on the child's sex). Using the Rasch-Welch DIF method (Linacre, 2020), I compared item difficulty estimates for males and females. DIF contrasts (i.e., systematic differences between item difficulty estimates) for males and females should be no larger than .43 logits (Zwick, Thayer & Lewis, 1999) to be considered negligible. I also conducted *t*-tests of item difficulty to examine the likelihood that DIF could be caused by chance alone. To ensure that tests meet the Rasch assumptions that easy items are easy for *all* individuals, items with both DIF contrast > .43 *and* $p < .05$ should be considered problematic and should be removed or targeted for revision. Given the large sample size in this study, I did not consider items to show bias if contrasts were significant but smaller than .43 logits.

**Internal Reliability**

I evaluated internal reliability based on two Winsteps-generated indices. The first, person reliability index is the Rasch equivalent to Cronbach's alpha and represents the amount of variance that can be reproduced by the Rasch model (Wright & Masters, 1982). Person reliability

index greater than 0.80 suggests strong evidence for internal reliability; greater than .70 is adequate (Bond, Yan & Heene, 2020).

The strata value is an additional measure of reliability that represents the number of levels of ability that the measure can distinguish (Wright & Masters, 1982). Winsteps generates a separation index (G), which I converted to strata using the formula:

$$Strata = \frac{4G + 1}{3}$$

Strata should be at least 2.0 to establish evidence for sufficient internal reliability (Bond, Yan & Heene, 2020). Given the developmental nature of sensory perception and the large age range of our sample, I expect higher strata values (i.e., I expect more levels of ability to be represented by the items). Therefore, I will consider strata values acceptable at 2.0 and strong at 3.0 or more.

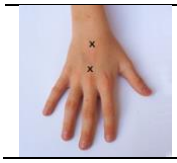
## Results

### Construct Validity

Tables 4.2-4.9 contain item measures, point-measure correlations, and fit statistics for each of the eight tests. All items showed positive point-measure correlations. Three items on TPD failed to fit (87.5% of items fit). Three items on VPS failed to fit (83.3% of items fit). One item on PF failed to fit (90.0% of items fit). PF, however, does reach our criteria for acceptable fit because only a single item misfit. All items across the five remaining tests fit the model.

**Table 4.2**

*Tactile Perception: Localization Item Measures, Fit Statistics and DIF Statistics*

| | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
|  | TPL 1 | 0.54 (0.05) | 1.00 | 1.01 | 0.47 | 0.09 |

| | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
|  | TPL 2 | -0.49 (0.06) | 1.00 | 1.03 | 0.40 | 0.00 |
|  | TPL 3 | 0.82 (0.05) | 1.02 | 1.05 | 0.47 | -0.08 |
|  | TPL 4 | -0.36 (0.06) | 1.02 | 1.00 | 0.40 | 0.07 |
|  | TPL 5 | -1.30 (0.08) | 1.00 | 1.11 | 0.34 | -0.18 |
|  | TPL 6 | 0.87 (0.05) | 0.94 | 0.92 | 0.52 | -0.12 |
|  | TPL 7 | 0.17 (0.05) | 1.02 | 1.05 | 0.43 | 0.05 |
|  | TPL 8 | -0.28 (0.06) | 1.05 | 1.02 | 0.39 | 0.12 |
|  | TPL 9 | 0.45 (0.05) | 0.98 | 1.02 | 0.46 | 0.00 |
|  | TPL 10 | 0.40 (0.05) | 1.07 | 1.08 | 0.42 | -0.18 |
|  | TPL 11 | 0.58 (0.05) | 0.93 | 0.92 | 0.50 | 0.15 |

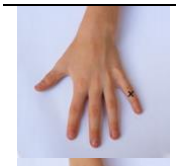
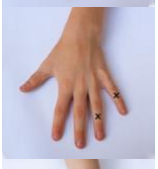
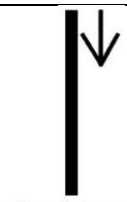
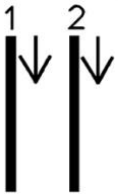
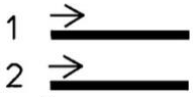
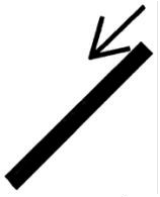
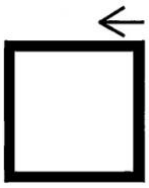
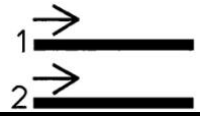
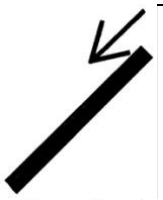
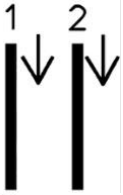
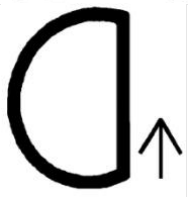
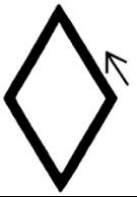
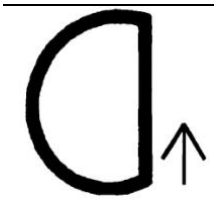
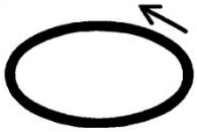
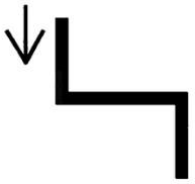
| | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
|  | TPL 12 | -0.92 (0.07) | 0.96 | 0.97 | 0.37 | 0.06 |
|  | TPL 13 | -0.5 (0.06) | 1.00 | 0.95 | 0.39 | -0.18 |
|  | TPL 14 | -0.48 (0.06) | 1.00 | 0.95 | 0.39 | -0.15 |
|  | TPL 15 | 0.87 (0.05) | 1.05 | 1.07 | 0.45 | -0.05 |
|  | TPL 16 | 0.42 (0.05) | 1.03 | 0.99 | 0.44 | 0.00 |
|  | TPL 17 | 1.05 (0.05) | 0.95 | 0.96 | 0.52 | 0.09 |
|  | TPL 18 | -1.72 (0.09) | 0.97 | 0.99 | 0.32 | -0.34 |
|  | TPL 19 | -0.37 (0.06) | 0.94 | 0.97 | 0.42 | 0.18 |
|  | TPL 20 | 0.24 (0.05) | 1.04 | 1.05 | 0.42 | 0.15 |

**Table 4.3**

*Tactile Perception: Designs Item Measures, Fit Statistics and DIF Statistics*

| | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
|  | TPD 1 | -2.46 (0.07) | 1.19 | **1.83** | 0.31 | -0.09 |
|  | TPD 2 | -0.99 (0.04) | 0.96 | 1.00 | 0.47 | 0.14 |
|  | TPD 3 | -1.24 (0.04) | 1.11 | **1.44** | 0.43 | 0.15 |
|  | TPD 4 | -1.13 (0.04) | 1.27 | **1.36** | 0.41 | -0.20* |
|  | TPD 5 | -0.98 (0.04) | 0.95 | 0.94 | 0.51 | 0.00 |
|  | TPD 6 | 0.05 (0.03) | 1.12 | 1.11 | 0.51 | 0.00 |
|  | TPD 7 | -0.15 (0.03) | 1.10 | 1.07 | 0.52 | 0.00 |
|  | TPD 8 | 0.37 (0.03) | 1.01 | 0.98 | 0.57 | -0.02 |
|  | TPD 9 | -0.96 (0.04) | 1.00 | 0.94 | 0.49 | -0.05 |

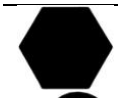
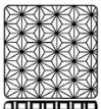
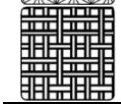
| | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
|  | TPD 10 | 0.06 (0.03) | 1.06 | 1.04 | 0.53 | -0.07 |
|   | TPD 11 | -1.07 (0.04) | 1.05 | 1.36 | 0.44 | -0.04 |
|  | TPD 12 | -0.52 (0.03) | 1.11 | 1.12 | 0.49 | 0.00 |
|  | TPD 13 | 0.63 (0.03) | 1.04 | 1.02 | 0.55 | -0.08 |
|   | TPD 14 | 0.41 (0.03) | 1.04 | 1.02 | 0.55 | 0.00 |
|  | TPD 15 | 1.54 (0.03) | 0.96 | 0.97 | 0.56 | -0.12 |
|  | TPD 16 | 0.34 (0.03) | 0.93 | 0.88 | 0.57 | 0.14* |
|  | TPD 17 | 1.44 (0.03) | 0.92 | 0.85 | 0.59 | -0.11 |

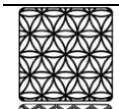
| | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
|  | TPD 18 | 0.38 (0.03) | 0.90 | 0.89 | 0.59 | 0.12* |
|  | TPD 19 | 0.41 (0.03) | 0.87 | 0.84 | 0.59 | 0.00 |
|  | TPD 20 | 0.14 (0.03) | 0.89 | 0.94 | 0.54 | 0.08 |
|  | TPD 21 | 1.26 (0.03) | 0.92 | 0.84 | 0.59 | 0.00 |
|  | TPD 22 | 0.73 (0.03) | 1.03 | 0.96 | 0.57 | -0.06 |
|  | TPD 23 | 1.32 (0.03) | 1.01 | 0.93 | 0.56 | 0.13* |
|  | TPD 24 | 0.43 (0.03) | 0.90 | 0.88 | 0.58 | 0.00 |

* = DIF statistics were significant at $p < .05$

**Table 4.4**

*Tactile Perception: Shapes Item Measures, Fit Statistics and DIF Statistics*

| | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
|  | TPS1.1 | 0.71 (0.05) | 1.11 | 1.15 | 0.41 | -0.06 |
|  | TPS1.2 | -1.8 (0.08) | 0.96 | 0.79 | 0.40 | -0.07 |
|  | TPS1.3 | 0.42 (0.05) | 1.04 | 1.02 | 0.45 | -0.02 |
|  | TPS1.4 | -0.31 (0.06) | 1.14 | 1.20 | 0.36 | -0.05 |
|  | TPS1.5 | -1.62 (0.08) | 0.94 | 0.83 | 0.41 | -0.15 |
|  | TPS1.6 | -0.7 (0.06) | 0.90 | 0.85 | 0.48 | -0.06 |
|  | TPS1.7 | -0.28 (0.06) | 0.92 | 0.90 | 0.48 | 0.00 |
|  | TPS1.8 | -0.22 (0.06) | 0.97 | 0.95 | 0.46 | -0.10 |
|  | TPS2.1 | -0.78 (0.06) | 0.96 | 0.91 | 0.44 | 0.11 |
|  | TPS2.2 | -1.34 (0.07) | 0.90 | 0.66 | 0.46 | -0.22 |
|  | TPS2.3 | -0.11 (0.05) | 0.98 | 0.90 | 0.46 | -0.20 |
|  | TPS2.4 | -1.27 (0.07) | 0.88 | 0.67 | 0.46 | 0.09 |
|  | TPS2.5 | -0.56 (0.06) | 0.86 | 0.79 | 0.50 | -0.32* |
|  | TPS2.6 | 1.12 (0.05) | 1.02 | 1.05 | 0.46 | 0.07 |
|  | TPS2.7 | -0.45 (0.06) | 0.93 | 0.84 | 0.46 | 0.00 |
|  | TPS2.8 | 0.9 (0.05) | 1.03 | 1.02 | 0.45 | -0.19 |
|  | TPS2.9 | 1.39 (0.05) | 1.10 | 1.16 | 0.41 | 0.00 |
|  | TPS2.10 | 1.62 (0.05) | 1.07 | 1.21 | 0.43 | 0.17 |

| | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
|  | TPS2.11 | 1.67 (0.05) | 1.07 | 1.12 | 0.43 | 0.05 |
|  | TPS2.12 | 1.62 (0.05) | 1.03 | 1.19 | 0.45 | 0.13 |

TPS1 = Tactile Perception Shapes Subtest 1; TPS2 = Tactile Perception Shapes Subtest 2; * = DIF statistics were significant at $p < .05$

**Table 4.5**

*Tactile Perception: Oral Item Measures, Fit Statistics and DIF Statistics*

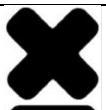
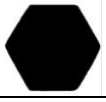
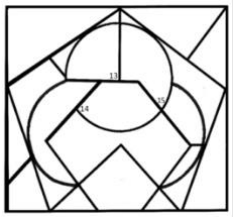
| | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
|  | TPO 1 | -0.62 (0.05) | 1.07 | 1.14 | 0.47 | -0.05 |
|  | TPO 2 | -0.20 (0.05) | 1.07 | 1.08 | 0.48 | -0.10 |
|  | TPO 3 | -0.88 (0.05) | 0.99 | 0.97 | 0.51 | 0.05 |
|  | TPO 4 | -0.41 (0.05) | 1.06 | 1.09 | 0.48 | -0.09 |
|  | TPO 5 | -0.02 (0.05) | 0.91 | 0.85 | 0.57 | -0.06 |
|  | TPO 6 | 0.98 (0.05) | 1.09 | 1.18 | 0.45 | 0.09 |
|  | TPO 7 | -0.62 (0.05) | 0.89 | 0.84 | 0.56 | -0.02 |
|  | TPO 8 | 0.85 (0.05) | 0.98 | 0.99 | 0.52 | 0.00 |
|  | TPO 9 | 0.63 (0.05) | 1.01 | 1.02 | 0.51 | 0.14 |
|  | TPO 10 | -0.33 (0.05) | 0.91 | 0.91 | 0.55 | 0.06 |
|  | TPO 11 | 0.63 (0.05) | 1.00 | 1.02 | 0.51 | 0.02 |

**Table 4.6**

*Visual Perception: Search Item Measures, Fit Statistics and DIF Statistics*

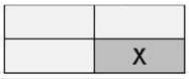
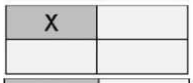
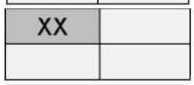
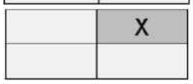
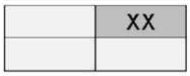
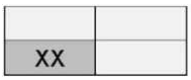
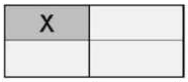
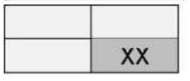
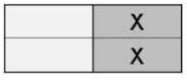
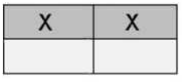
| | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
|  | VPST 1 | -3.98 (0.19) | 0.96 | 0.71 | 0.27 | 0.23 |
| | VPST 2 | -3.25 (0.14) | 0.96 | 0.75 | 0.31 | **-0.57** |
| | VPST 3 | -3.38 (0.15) | 0.94 | 1.96 | 0.29 | -0.16 |
| | VPST 4 | -2.30 (0.1) | 0.99 | 0.74 | 0.37 | -0.16 |
| | VPST 5 | -0.71 (0.07) | 1.16 | **1.41** | 0.38 | 0.19 |
| | VPST 6 | -0.64 (0.06) | 1.16 | **1.59** | 0.38 | 0.00 |
|  | VPSS 1 | -0.23 (0.06) | 1.01 | 0.94 | 0.49 | 0.24* |
| | VPSS 2 | -0.75 (0.07) | 0.88 | 0.78 | 0.51 | 0.40* |
| | VPSS 3 | 0.75 (0.05) | 0.97 | 0.95 | 0.54 | 0.00 |
| | VPSS 4 | 0.76 (0.05) | 1.00 | 1.10 | 0.52 | 0.00 |
| | VPSS 5 | 1.77 (0.05) | 1.01 | 1.03 | 0.54 | 0.00 |
| | VPSS 6 | 1.5 (0.05) | 1.10 | 1.14 | 0.49 | 0.00 |
|  | VPSL 1 | 0.57 (0.05) | 0.88 | 0.80 | 0.58 | -0.09 |
| | VPSL 2 | 0.74 (0.05) | 0.96 | 0.95 | 0.54 | 0.15 |
| | VPSL 3 | 1.3 (0.05) | 0.87 | 0.80 | 0.60 | -0.05 |
| | VPSL 4 | 2.78 (0.05) | 0.94 | 1.02 | 0.56 | -0.21* |
| | VPSL 5 | 2.8 (0.05) | 0.99 | **1.55** | 0.53 | 0.06 |
| | VPSL 6 | 2.28 (0.05) | 1.07 | 1.30 | 0.50 | -0.28* |

VPST = Visual Perception Search Toys Form; VPSS = Visual Perception Search Shells Form; VPSL = Visual Perception Search Lines Form, * = DIF statistics were significant at $p < .05$

**Table 4.7**

*Auditory Localization Item Measures, Fit Statistics and DIF Statistics*

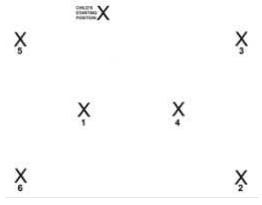
| Item Description | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
| Above head | ALC 1 | -0.89 (0.05) | 1.04 | 1.05 | 0.33 | 0.05 |
| Left shoulder | ALC 2 | -1.91 (0.07) | 0.95 | 0.82 | 0.33 | 0.15 |
| Left hip | ALC 3 | 0.32 (0.05) | 0.97 | 0.97 | 0.45 | 0.00 |
| Right shoulder | ALC 4 | -1.83 (0.07) | 0.97 | 0.96 | 0.31 | 0.16 |
| Above head | ALC 5 | -0.65 (0.05) | 1.01 | 1.01 | 0.37 | 0.07 |
| Left shoulder, then left hip | ALC 6 | 0.51 (0.05) | 0.90 | 0.86 | 0.51 | -0.09 |
| Left shoulder, then right hip | ALC 7 | 0.54 (0.05) | 0.90 | 0.85 | 0.51 | 0.00 |

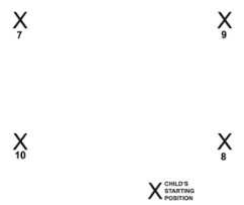
| Item Description | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
| Right hip, then left shoulder | ALC 8 | 0.40 (0.05) | 0.91 | 0.90 | 0.49 | -0.17 |
|  | ALT 1 | -0.74 (0.05) | 1.13 | 1.18 | 0.27 | 0.26* |
|  | ALT 2 | -0.20 (0.05) | 0.97 | 0.92 | 0.43 | -0.06 |
|  | ALT 3 | 0.17 (0.05) | 0.97 | 0.98 | 0.44 | 0.12 |
|  | ALT 4 | -0.26 (0.05) | 0.96 | 0.96 | 0.42 | -0.02 |
|  | ALT 5 | 0.16 (0.05) | 1.02 | 1.03 | 0.40 | -0.11 |
|  | ALT 6 | 0.20 (0.05) | 1.09 | 1.10 | 0.36 | 0.00 |
|  | ALT 7 | 1.24 (0.05) | 1.04 | 1.04 | 0.43 | -0.11 |
|  | ALT 8 | -0.18 (0.05) | 0.95 | 0.89 | 0.44 | 0.00 |
|  | ALT 9 | 0.16 (0.05) | 1.11 | 1.15 | 0.33 | 0.25* |
|  | ALT 10 | 0.79 (0.05) | 1.05 | 1.06 | 0.41 | 0.00 |
|  | ALT 11 | 0.77 (0.05) | 1.03 | 1.04 | 0.42 | -0.10 |
|  | ALT 12 | 1.41 (0.05) | 1.04 | 1.07 | 0.43 | -0.17 |

ALC = Auditory Localization Clicker; ALT = Auditory Localization Table, * = DIF statistics were significant at $p < .05$

**Table 4.8**

*Proprioception: Joint Position Item Measures, Fit Statistics and DIF Statistics*

| Item Description | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
|  | PJPH 1 | -0.19 (0.03) | 1.00 | 0.99 | 0.44 | 0.00 |
| | PJPH 2 | 0.34 (0.03) | 1.07 | 1.08 | 0.44 | 0.00 |
| | PJPH 3 | -0.03 (0.03) | 0.99 | 1.00 | 0.43 | 0.06 |
| | PJPH 4 | -0.06 (0.03) | 0.98 | 0.96 | 0.48 | 0.18* |
| | PJPF 5 | -0.42 (0.03) | 0.96 | 0.95 | 0.45 | -0.06 |
| | PJPF 6 | 0.14 (0.03) | 1.01 | 1.02 | 0.43 | 0.05 |

| Item Description | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
|  | PJPF 1 | 0.86 (0.03) | 1.02 | 1.02 | 0.45 | 0.00 |
| | PJPF2 | 0.58 (0.03) | 0.99 | 0.98 | 0.51 | 0.00 |
| | PJPF 3 | 0.64 (0.03) | 1.00 | 0.99 | 0.47 | 0.12 |
| | PJPF 4 | 0.75 (0.03) | 1.05 | 1.06 | 0.46 | -0.06 |
| See note | PJP2H 1 | -0.51 (0.04) | 0.96 | 0.96 | 0.45 | -0.02 |
| | PJP2H 2 | -0.53 (0.04) | 0.98 | 0.98 | 0.45 | 0.04 |
| | PJP2H 3 | -0.58 (0.04) | 0.97 | 0.97 | 0.46 | -0.11 |
| | PJP2H 4 | -0.53 (0.04) | 1.00 | 1.01 | 0.44 | -0.10 |
| | PJP2H 5 | -0.47 (0.04) | 0.96 | 0.97 | 0.44 | -0.15* |

Two hand items are different "zones" along the vertical surface of a door, approximately 2", with zone 1 at the base of the door and each section 2" higher; PJPH = Proprioception Joint Positions Hand Items; PJPF = Proprioception Joint Positions Foot Items; PJP2H = Proprioception Joint Positions Two Hand Items, * = DIF statistics were significant at $p < .05$

**Table 4.9**

*Proprioception: Force Item Measures, Fit Statistics and DIF Statistics*

| Item Description | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
| Dark circle, preferred hand | PFC 1 | -1.26 (0.05) | 0.91 | 0.71 | 0.49 | 0.00 |
| Dark circle, nonpreferred hand | PFC 2 | 0.09 (0.03) | 0.82 | 0.80 | 0.51 | -0.06 |
| Mid-range circle, preferred hand | PFC 3 | -0.84 (0.04) | 0.91 | 0.81 | 0.48 | 0.00 |
| Mid-range circle, nonpreferred hand | PFC 4 | 0.06 (0.03) | 0.79 | 0.80 | 0.53 | 0.16 |
| Light circle, preferred hand | PFC 5 | -0.70 (0.04) | 1.00 | 0.88 | 0.49 | -0.10 |
| Light circle, nonpreferred hand | PFC 6 | -0.38 (0.04) | 0.92 | 0.87 | 0.48 | 0.00 |
| Rolling bottle, preferred hand | PFB 1 | 0.62 (0.03) | 1.16 | 1.17 | 0.52 | 0.00 |
| Rolling bottle, nonpreferred hand | PFB 2 | 0.65 (0.03) | 1.17 | 1.21 | 0.56 | 0.00 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Rolling bottle, both hands | PFB 3 | 0.55 (0.03) | 1.09 | 1.09 | 0.53 | 0.10 |
| Rolling bottle, both feet | PFB 4 | 1.20 (0.03) | 1.25 | **1.31** | 0.53 | -0.16* |

PFC = Proprioception Force Crayon Items; PFB = Proprioception Force Bottle Items, * = DIF statistics were significant at $p < .05$

Table 4.10 displays the results of the person fit analyses. Four of the tests (TPD, TPS, VPS, and PF) fell below our criteria. Notably, the tests with more misfit still had at least 84.8% person fit.

**Table 4.10**

*Person Fit Analysis*

| Test | Children with Misfitting Data | Total Number of Children | % fitting children |
|---|---|---|---|
| Tactile Perception: Localization | 167 | 2394 | 93.0% |
| Tactile Perception: Designs | 362 | 2383 | 84.8% |
| Tactile Perception: Shapes | 253 | 2316 | 89.1% |
| Tactile Perception: Oral | 163 | 2262 | 92.8% |
| Visual Perception: Shapes | 347 | 2498 | 86.1% |
| Auditory Localization | 169 | 2410 | 93.0% |
| Proprioception: Joint Position | 178 | 2367 | 92.5% |
| Proprioception: Force | 347 | 2446 | 85.8% |

Table 4.11 contains the results of rating scale analyses. Across all tests, fit statistics aligned to the expectations of the Rasch model. All polytomous rating scales demonstrated acceptable Andrich thresholds. Observed average person measures increased monotonically for each rating scale category.

**Table 4.11**

*Rating Scale Analysis for Sensory Perception Scales*

| Item Type | Rating Scale Category | % Used | Infit MnSq | Outfit MnSq | Andrich Threshold[1] | Observed Average |
|---|---|---|---|---|---|---|
| Tactile Perception: Localization | | | | | | |
| Accuracy | 0 | 23.08% | 1.00 | 1.00 | – | .40 |
| | 1 | 76.92% | 1.00 | 1.01 | – | 1.69 |
| Tactile Perception: Designs | | | | | | |
| Accuracy | 0 | 28.18% | 1.01 | 1.17 | – | -0.70 |
| | 1 | 19.24% | 0.95 | 0.96 | 0.17 | 0.31 |
| | 2 | 52.58% | 1.01 | 0.98 | -0.17 | 1.38 |
| Tactile Perception: Shapes | | | | | | |
| Accuracy | 0 | 29.15% | 0.99 | 0.89 | – | -0.08 |
| | 1 | 70.85% | 1.03 | 1.11 | – | 1.76 |
| Tactile Perception: Oral | | | | | | |
| Accuracy | 0 | 43.59% | 1.00 | 0.99 | – | -0.44 |
| | 1 | 56.41% | 1.00 | 1.03 | – | 0.89 |
| Visual Perception: Search | | | | | | |
| Accuracy | 0 | 29.45% | 1.00 | 1.06 | – | -0.66 |
| | 1 | 70.55% | 0.99 | 1.14 | – | 2.58 |
| Auditory Localization | | | | | | |
| Accuracy | 0 | 37.70% | 1.00 | .99 | – | -0.18 |
| | 1 | 62.30% | 1.00 | 1.00 | – | 1.12 |
| Proprioception: Joint Position | | | | | | |
| Accuracy (Hand/Foot Distance) | 0 | 21.02% | 0.94 | 0.92 | – | -0.52 |
| | 1 | 39.98% | 1.03 | 1.04 | -0.73 | 0.37 |
| | 2 | 39.00% | 1.04 | 1.05 | 0.73 | 1.06 |
| Accuracy (Two Hand Distance) | 0 | 6.94% | 0.93 | 0.91 | – | 0.19 |
| | 1 | 45.98% | 0.99 | 1.01 | -1.24 | 0.99 |
| | 2 | 47.08% | 1.00 | 1.00 | 1.24 | 1.53 |
| Proprioception: Force | | | | | | |
| Accuracy (Crayon) | 0 | 7.56% | 0.90 | 0.79 | – | .16 |
| | 1 | 25.45% | 0.88 | 0.76 | -0.53 | .96 |
| | 2 | 66.99% | 0.86 | 0.91 | 0.53 | 1.93 |
| Accuracy (Bottle) | 0 | 24.14% | 1.11 | 1.09 | – | -.31 |
| | 1 | 38.19% | 1.13 | 1.17 | -0.58 | .24 |
| | 2 | 37.67% | 1.23 | 1.32 | 0.58 | 1.93 |

Principal components analysis revealed possible additional dimensions for PJP and PF, with eigenvalues greater than 2.0 (see Table 4.13 for details). Contrast loadings for PJP suggest that the Two Hand items diverge from the Hand/Foot items (see Table 4.14; disattenuated correlation 0.20). For PF, the Crayon items diverge from the Bottle items (see Table 4.15; disattenuated correlation 0.27). The remaining tests revealed no evidence for additional dimensions.

**Table 4.13**

Principal Components Analysis of Standardized Rasch Residuals

| Test | Eigenvalue of Largest Contrast | Variance Explained by Largest Contrast | Variance Explained by Rasch Dimension |
|---|---|---|---|
| Tactile Perception: Localization | 1.89 | 7.4% | 21.6% |
| Tactile Perception: Designs | 1.63 | 3.7% | 45.4% |
| Tactile Perception: Shapes | 1.55 | 5.4% | 30.0% |
| Tactile Perception: Oral | 1.35 | 9.2% | 25.8% |
| Visual Perception: Search | 1.63 | 5.1% | 43.7% |
| Auditory Localization | 1.98 | 7.5% | 23.8% |
| Proprioception: Joint Position | 2.91 | 14.4% | 30.3% |
| Proprioception: Force | 2.11 | 14.2% | 32.7% |

**Table 4.14**

PJP Contrast Loadings

| Item | Loading |
|---|---|
| Two Hand Positions 4 | 0.69 |
| Two Hand Positions 3 | 0.67 |
| Two Hand Positions 2 | 0.66 |
| Two Hand Positions 5 | 0.62 |
| Two Hand Positions 1 | 0.61 |
| Hand Positions 6 | -0.22 |
| Hand Positions 1 | -0.22 |
| Hand Positions 3 | -0.23 |
| Hand Positions 5 | -0.24 |
| Hand Positions 4 | -0.28 |
| Foot Positions 3 | -0.29 |
| Foot Positions 2 | -0.31 |

| Item | Loading |
|---|---|
| Foot Positions 4 | -0.33 |
| Foot Positions 1 | -0.33 |
| Hand Positions 2 | -0.33 |

**Table 4.15**

PF Contrast Loadings

| Item | Loading |
|---|---|
| Crayon 2 | 0.53 |
| Crayon 4 | 0.48 |
| Crayon 1 | 0.45 |
| Crayon 3 | 0.41 |
| Crayon 5 | 0.40 |
| Crayon 6 | 0.40 |
| Bottle 1 | -0.49 |
| Bottle 2 | -0.47 |
| Bottle 3 | -0.47 |
| Bottle 4 | -0.47 |

DIF analysis revealed no items with significant and large DIF based on sex (see Tables 4.2-4.9). One item on VPS (Form 1.2) had a large contrast size but did not reach significance.

Figures 4.1-4.8 show the Winsteps-generated Wright maps for each of the sensory perception tests. For TPL, person abilities greatly exceed item ability, with children clustered at the top of the map and most items falling well below the mean person ability level. I observed the same pattern on TPS, TPO and VPS, PJP and PF, although to a lesser extent. These observations align with the mean person abilities shown in Table 4.16; while mean item measures are fixed at zero, the item difficulties are all above zero. Notably, the mean and standard deviations reported on these figures omit children with maximum and minimum scores, as these children's measures cannot be estimated in the Rasch model. Table 4.12 contains the mean measure scores including *all* participating children.

On most tests, I observed a range of item difficulties capturing multiple levels of person ability. However, on TPD, six items had nearly the same item measure (Designs 14, 16, 18, 19, 24, and 8). Other tests demonstrated redundancy as well, with most tests showing two to three overlapping items. Additionally, I observed difficulty compression on both PF and PJP; items spanned difficulties approximately +1 to -1 logits, while child abilities spanned much larger ranges.

The item hierarchies appeared logical. On TPL, items that required two-point discrimination were more challenging than items that required only single-point localization. On TPD, designs requiring changes in direction (e.g., shapes and crosses) were more challenging than those with only single or multiple straight lines. On TPS, texture items were more difficult than shapes. On TPO, shapes with few discriminating features (i.e., hexagon [6], octagon [11], and pentagon [8]) were more challenging than shapes with clear points (i.e., asterisk [3], X shape [1], Y shape [11]). On VPS, Forms 1 through 6 become progressively more challenging; this is logical, as the first forms show brightly and differently colored objects, while the later forms are white line drawings that are more difficult to identify in the time frame. On AL, localizing sounds was more difficult on a table than on the child's body. Localizing two sounds was more difficult than localizing a single sound. On PJP, recreating hand positions was simpler than recreating foot positions. Additionally, matching two limbs at the same time was easier than returning a single limb to a position placed by the examiner. On PF, the Crayon items, in which the child received constant visual feedback on the darkness of their drawings, were easier than the Bottle items, in which the child had only one opportunity to grade force. Furthermore, drawing with a crayon is a familiar task, while rolling a rice bottle is a novel task.

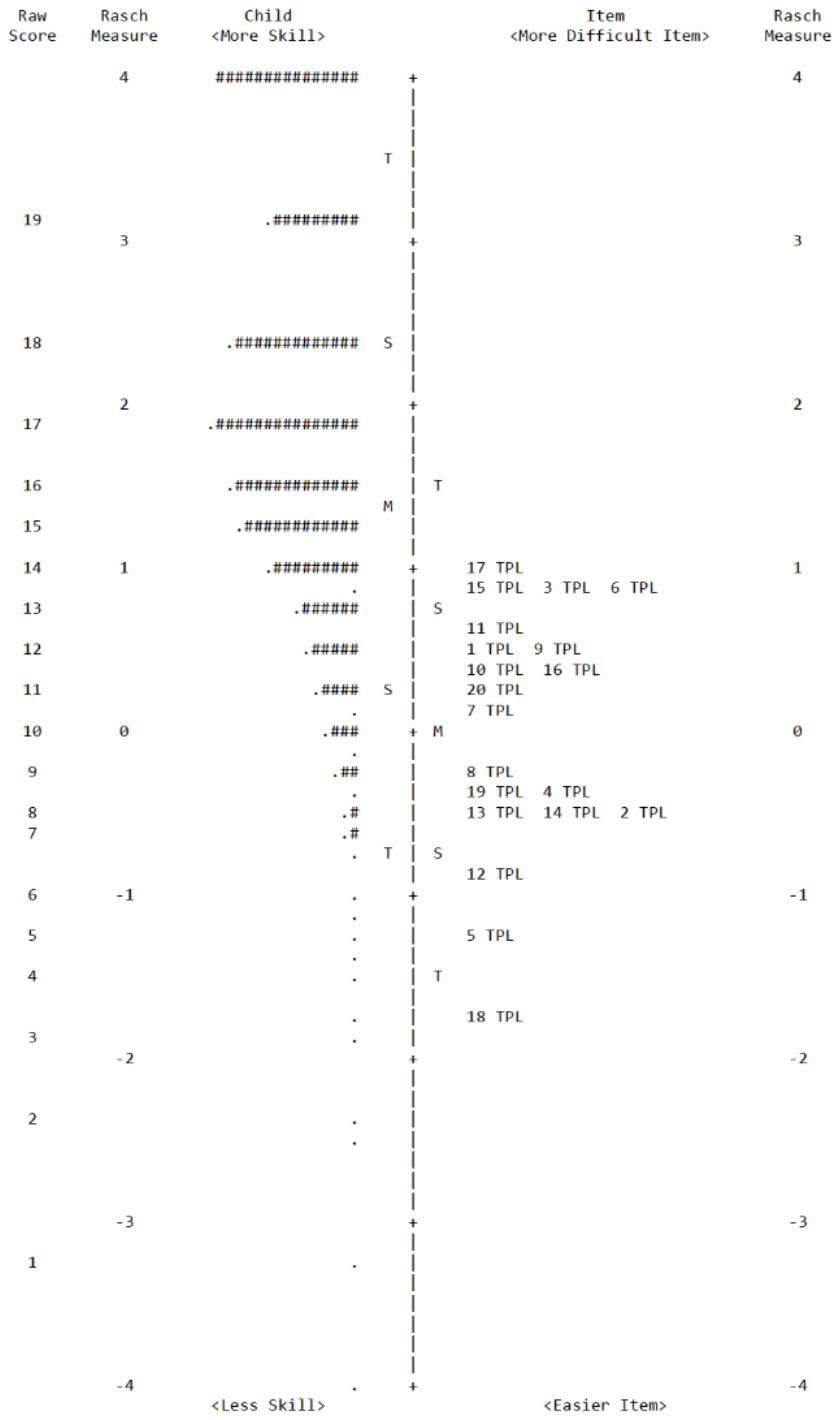



**Figure 4.1**

*Tactile Perception: Localization Wright Map*

*Note*. # = 20 children,  = 1 to 19 children, | = latent trait (tactile perception), M = item/child mean, S = 1 SD, T = 2 SD

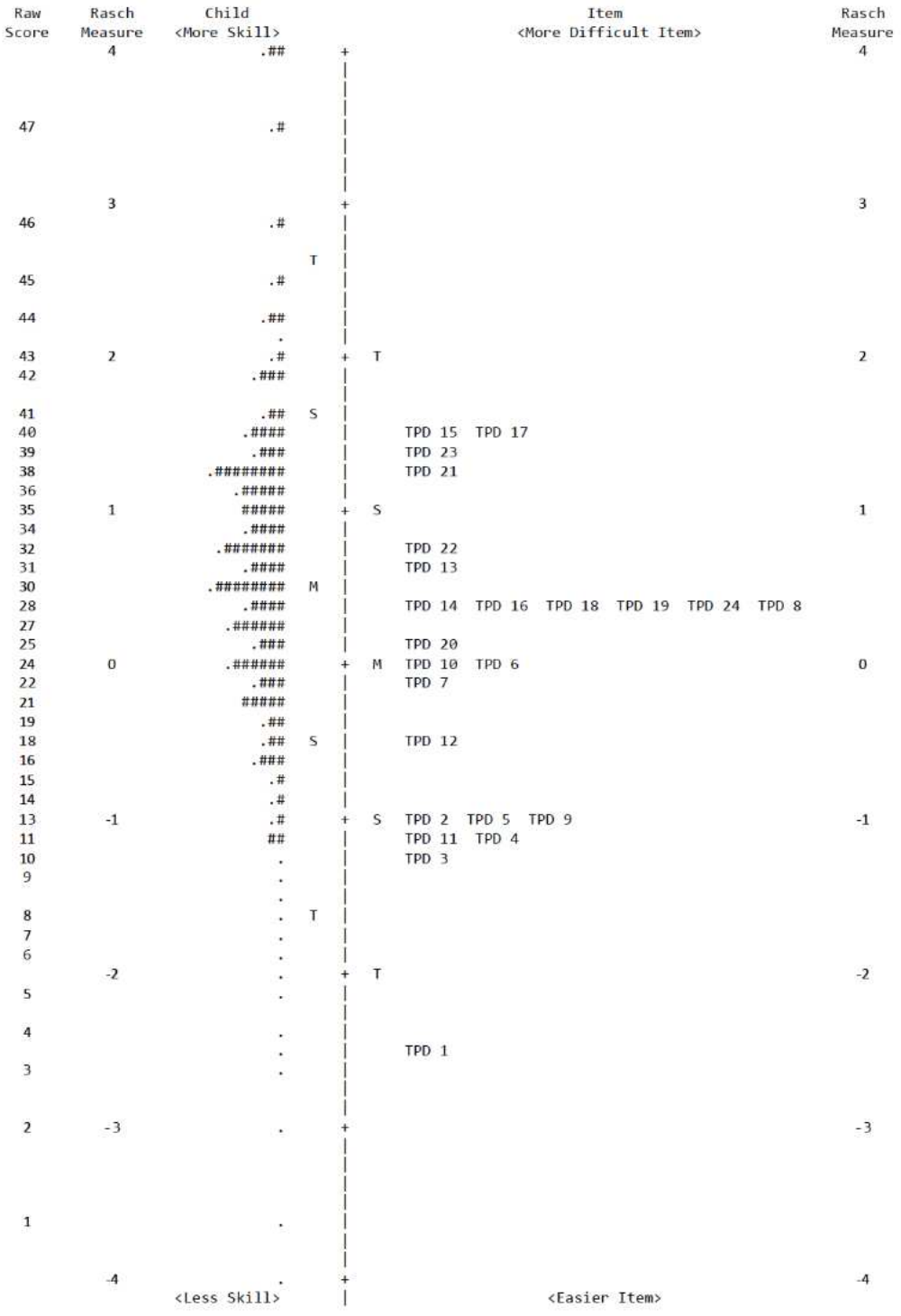



**Figure 4.2**

*Tactile Perception: Designs Wright Map*

*Note*. # = 20 children, . = 1 to 19 children, | = latent trait (tactile perception), M = item/child mean, S = 1 SD, T = 2 SD

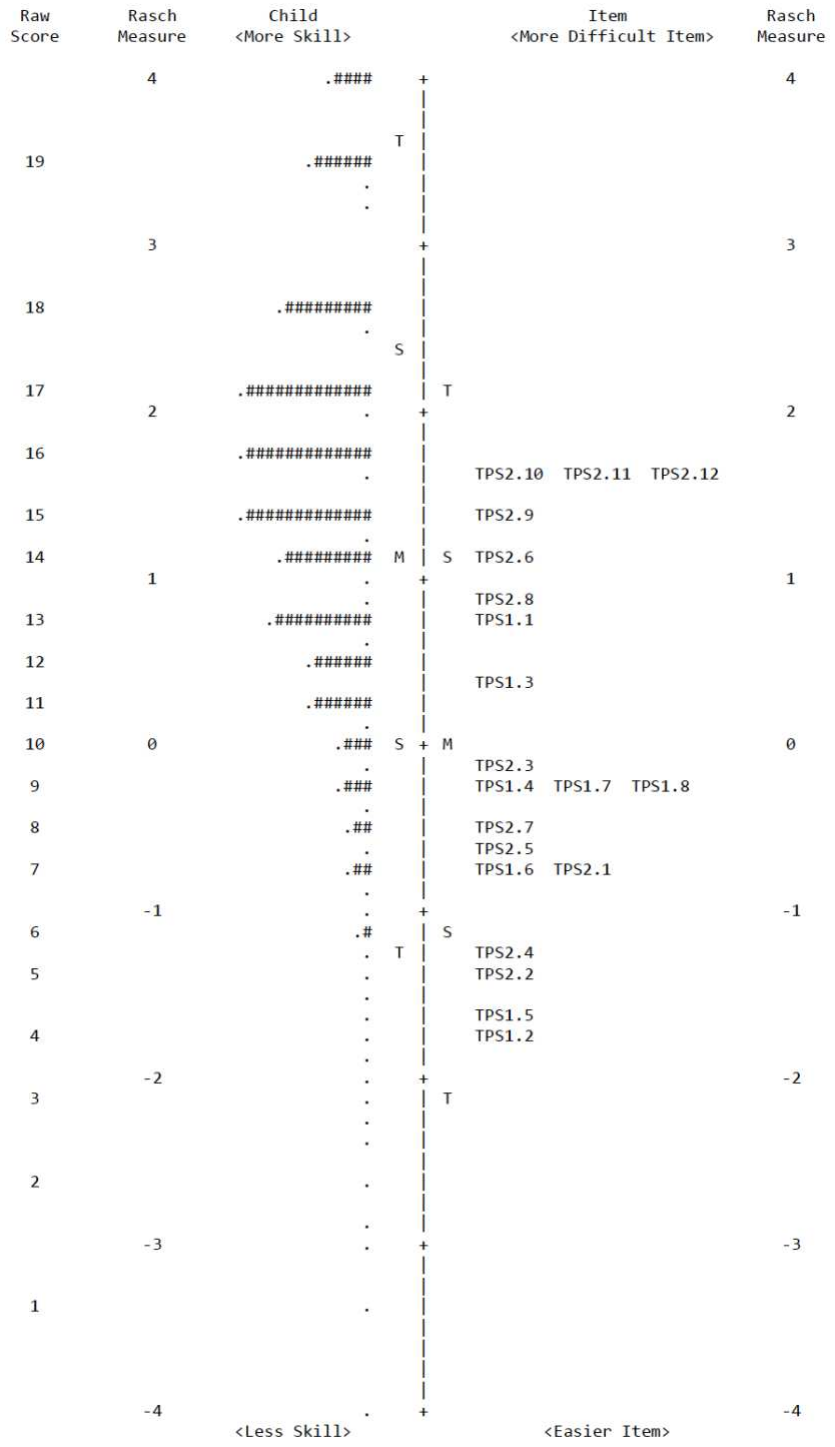



**Figure 4.3**

*Tactile Perception: Shapes Wright Map*

*Note*. # = 20 children, . = 1 to 19 children, | = latent trait (tactile perception), M = item/child mean, S = 1 SD, T = 2 SD

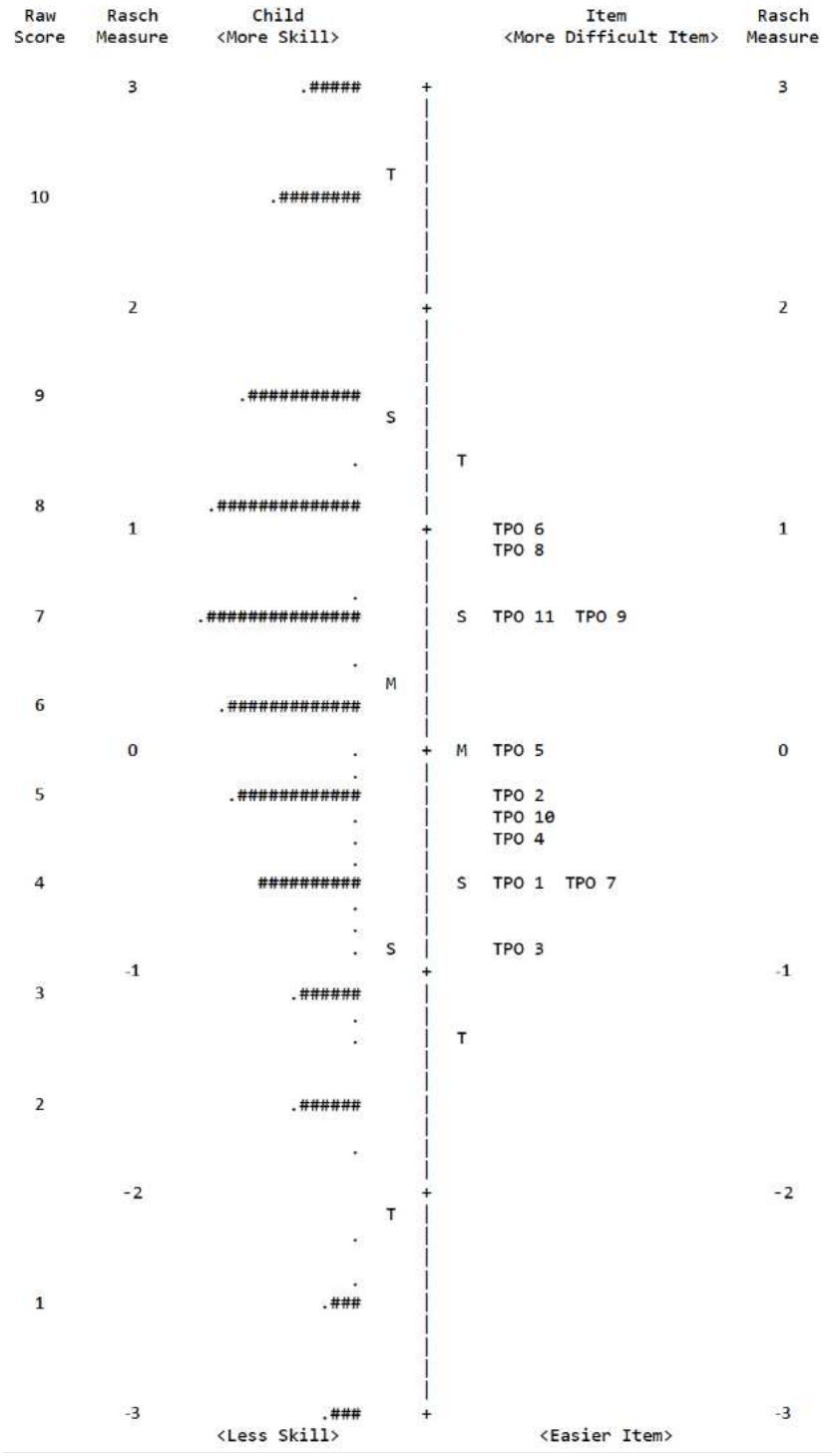



**Figure 4.4**

*Tactile Perception: Oral Wright Map*

*Note*. # = 20 children, . = 1 to 19 children, | = latent trait (tactile perception), M = item/child mean, S = 1 SD, T = 2 SD

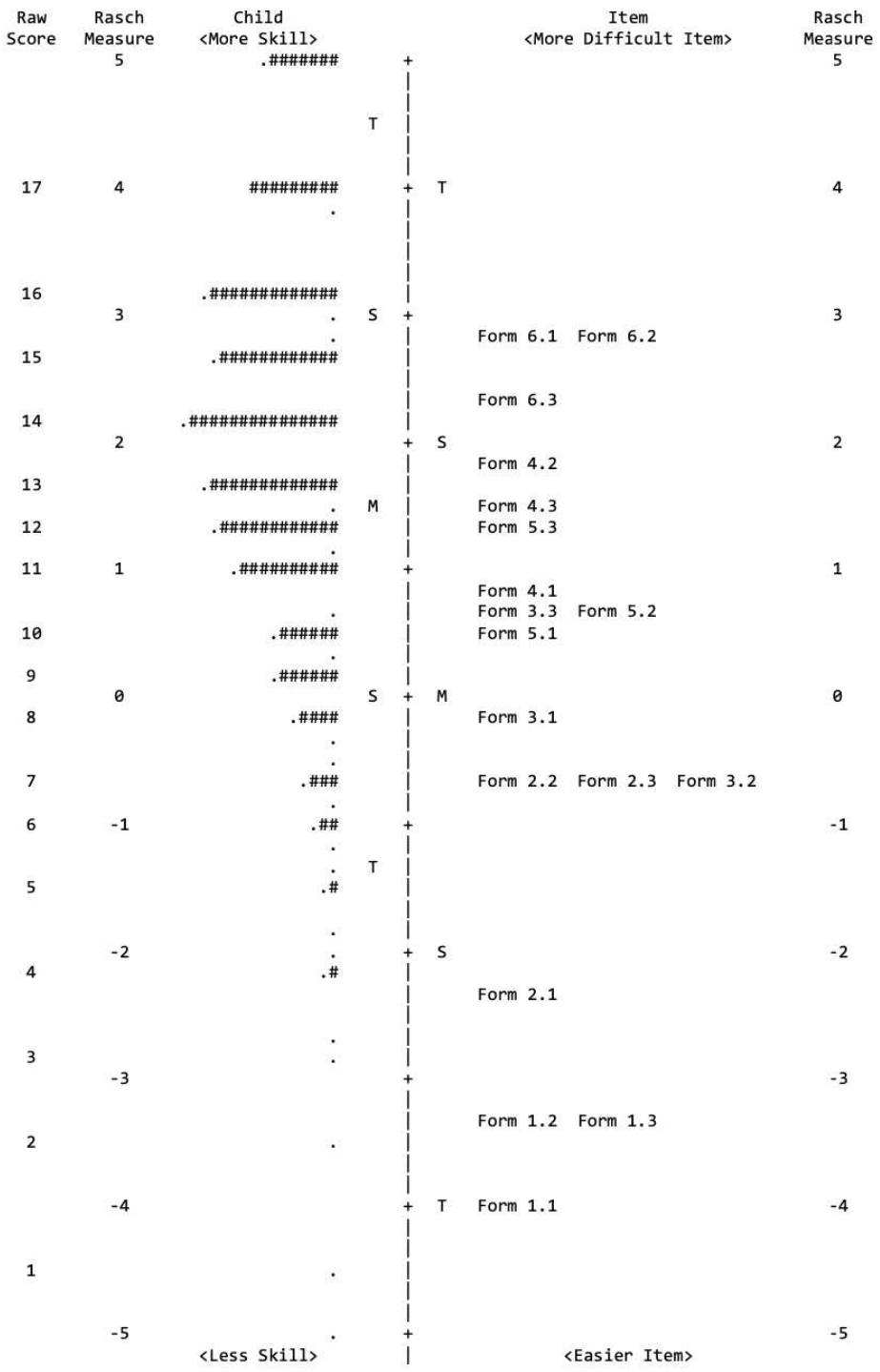



**Figure 4.5**

*Visual Perception: Search Wright Map*

*Note*. # = 20 children, . = 1 to 19 children, | = latent trait (visual perception), M = item/child mean, S = 1 SD, T = 2 SD

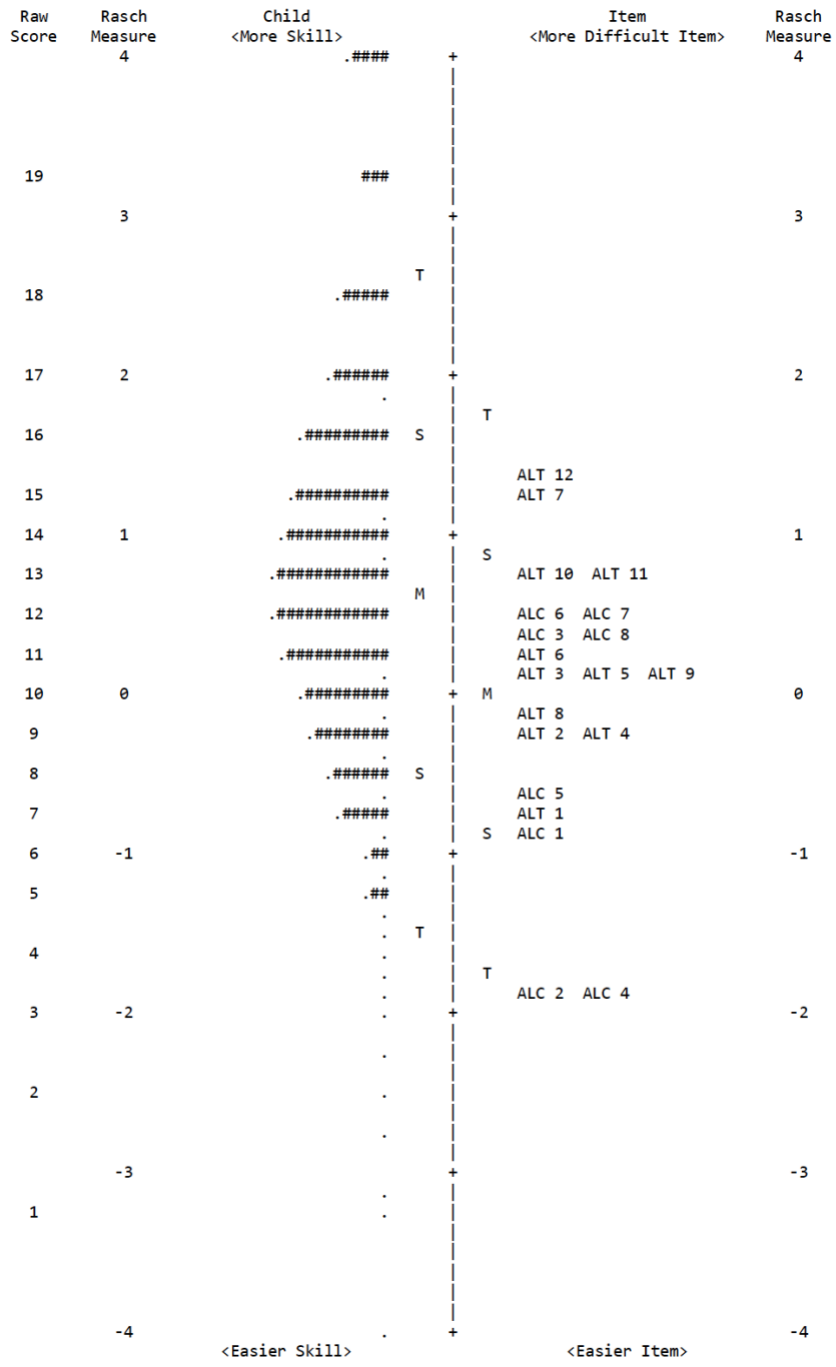



**Figure 4.6**

*Auditory Localization Wright Map*

*Note*. # = 20 children, . = 1 to 19 children, | = latent trait (auditory localization), M = item/child mean, S = 1 SD, T = 2 SD

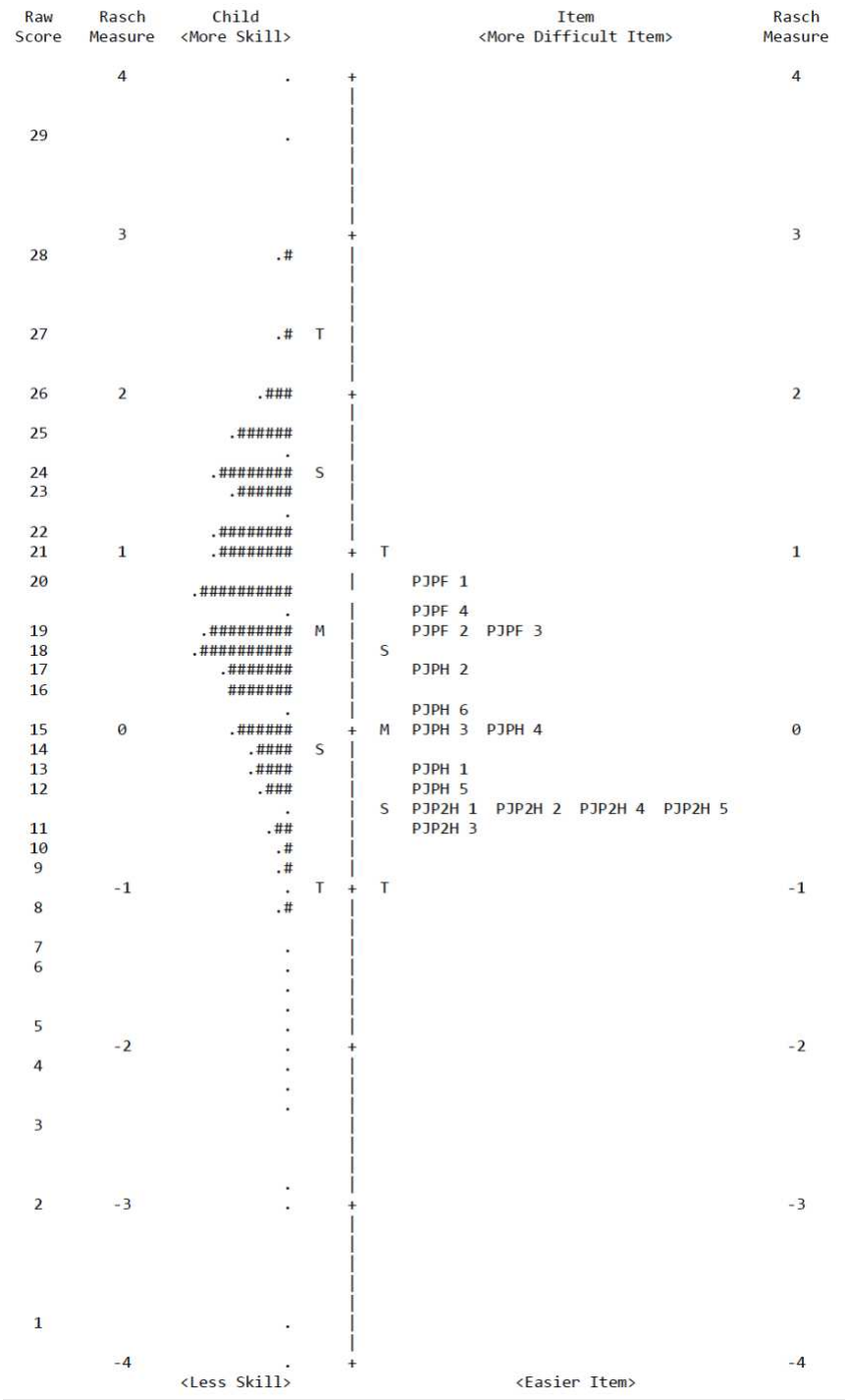



**Figure 4.7**

*Proprioception: Joint Position Wright Map*

*Note*. # = 20 children, . = 1 to 19 children, | = latent trait (proprioception), M = item/child mean, S = 1 SD, T = 2 SD

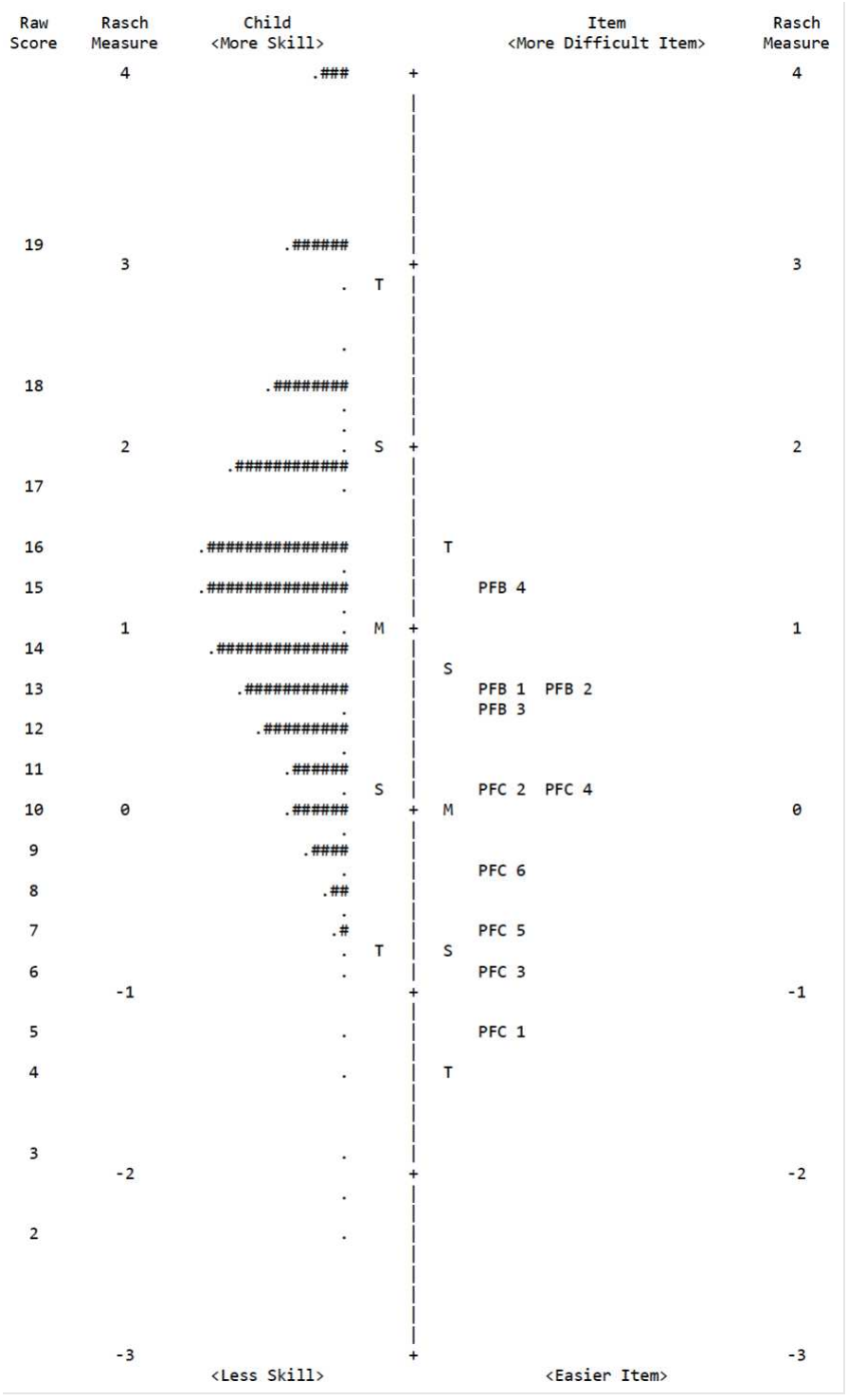



**Figure 4.8**

*Proprioception: Force Wright Map*

*Note*. # = 20 children, . = 1 to 19 children, | = latent trait (proprioception), M = item/child mean, S = 1 SD, T = 2 SD

**Table 4.16**

*Mean Child Measure Scores*

| Test | Child Mean (SE) |
|---|---|
| Tactile Perception: Localization | 1.69 (0.78) |
| Tactile Perception: Designs | 0.59 (0.37) |
| Tactile Perception: Shapes | 1.28 (0.68) |
| Tactile Perception: Oral | 0.33 (0.82) |
| Visual Perception: Search | 1.75 (0.77) |
| Auditory Localization | 0.73 (0.60) |
| Proprioception: Joint Position | 0.67 (0.44) |
| Proprioception: Force | 1.11 (0.61) |

Figures 4.9-4.16 show the person maps for the sensory perception tests. The person maps demonstrated expected age progression in person measure at each level (i.e., older children tended to score higher on all sensory perception tests). Table 4.17 contains the results of Pearson correlations between Rasch-generated measure scores and age in months.

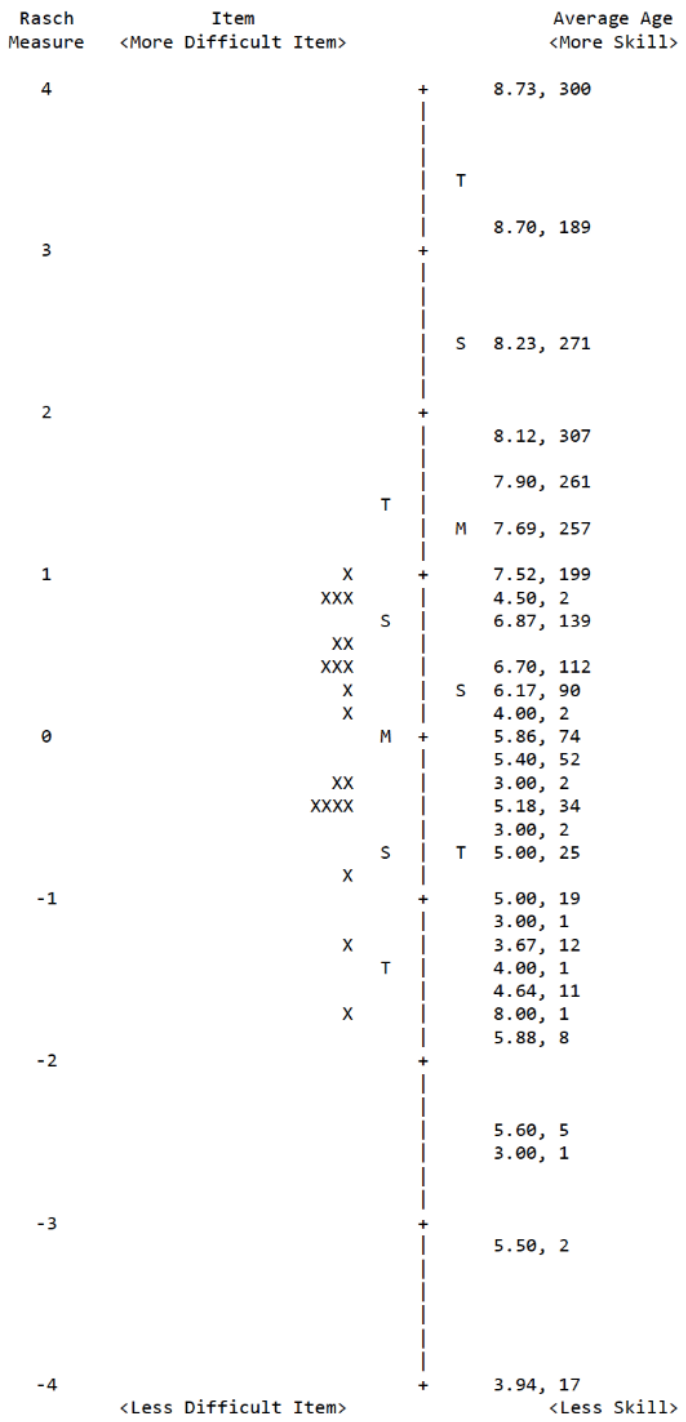



**Figure 4.9**

*Tactile Perception: Localization Person Map*

*Note.* | = latent trait (tactile perception), left side of the figure shows items (X), right side of the figure shows (average age, # of children), M = item/child mean, S = 1 SD, T = 2 SD

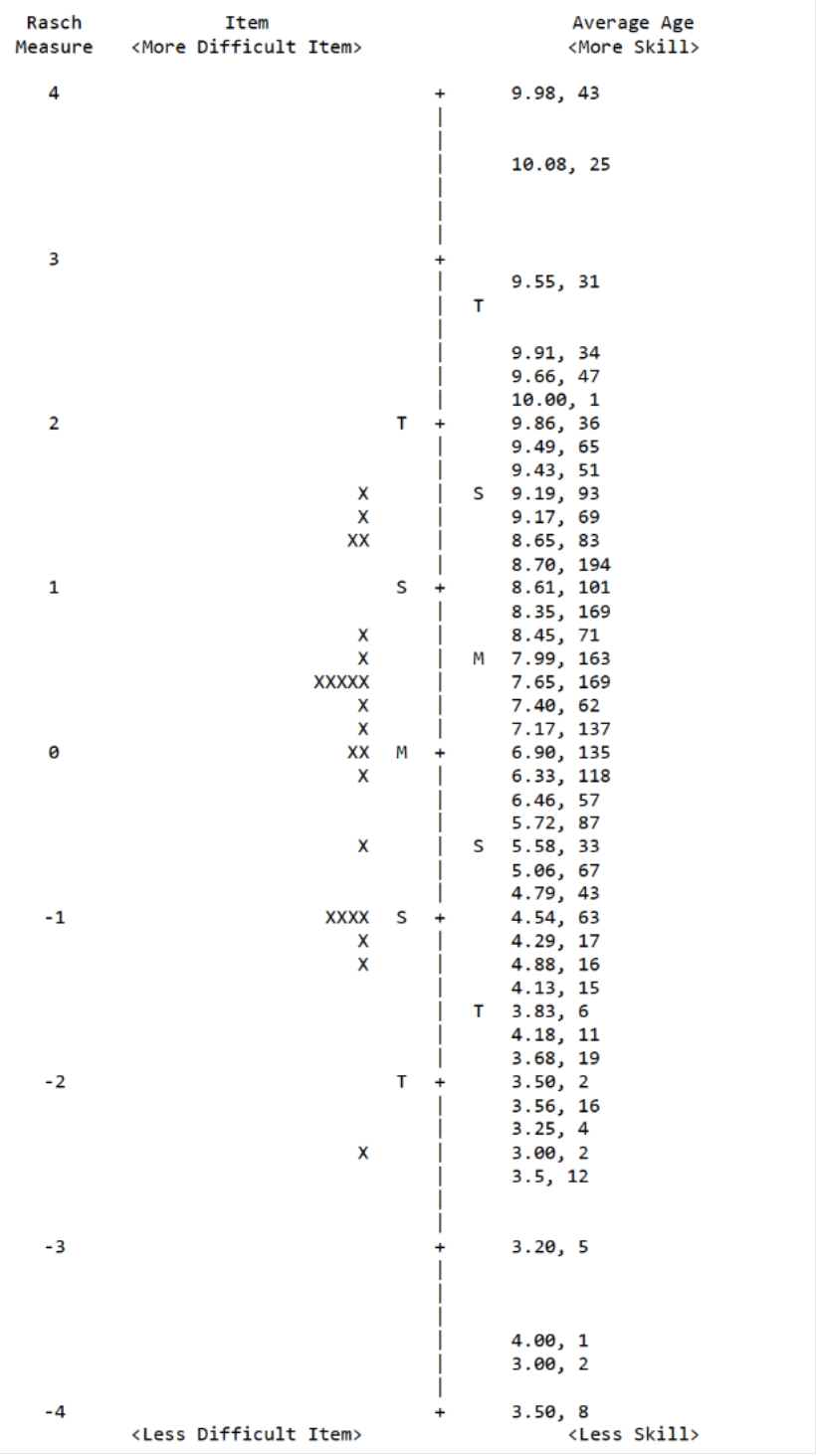



**Figure 4.10**

*Tactile Perception: Designs Person Map*

*Note.* | = latent trait (tactile perception), left side of the figure shows items (X), right side of the figure shows (average age, # of children), M = item/child mean, S = 1 SD, T = 2 SD

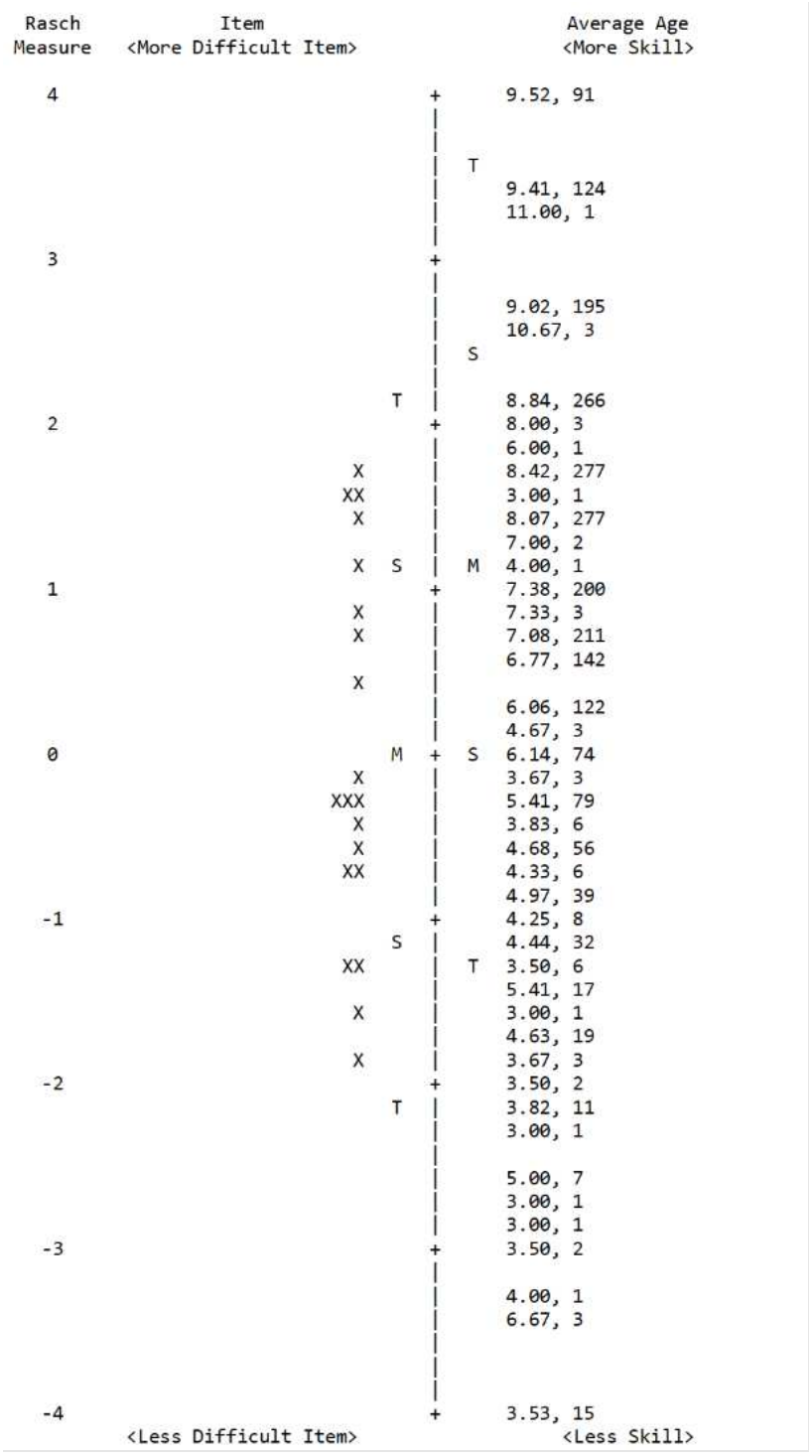



**Figure 4.11**

*Tactile Perception: Shapes Person Map*

*Note.* | = latent trait (tactile perception), left side of the figure shows items (X), right side of the figure shows (average age, # of children), M = item/child mean, S = 1 SD, T = 2 SD

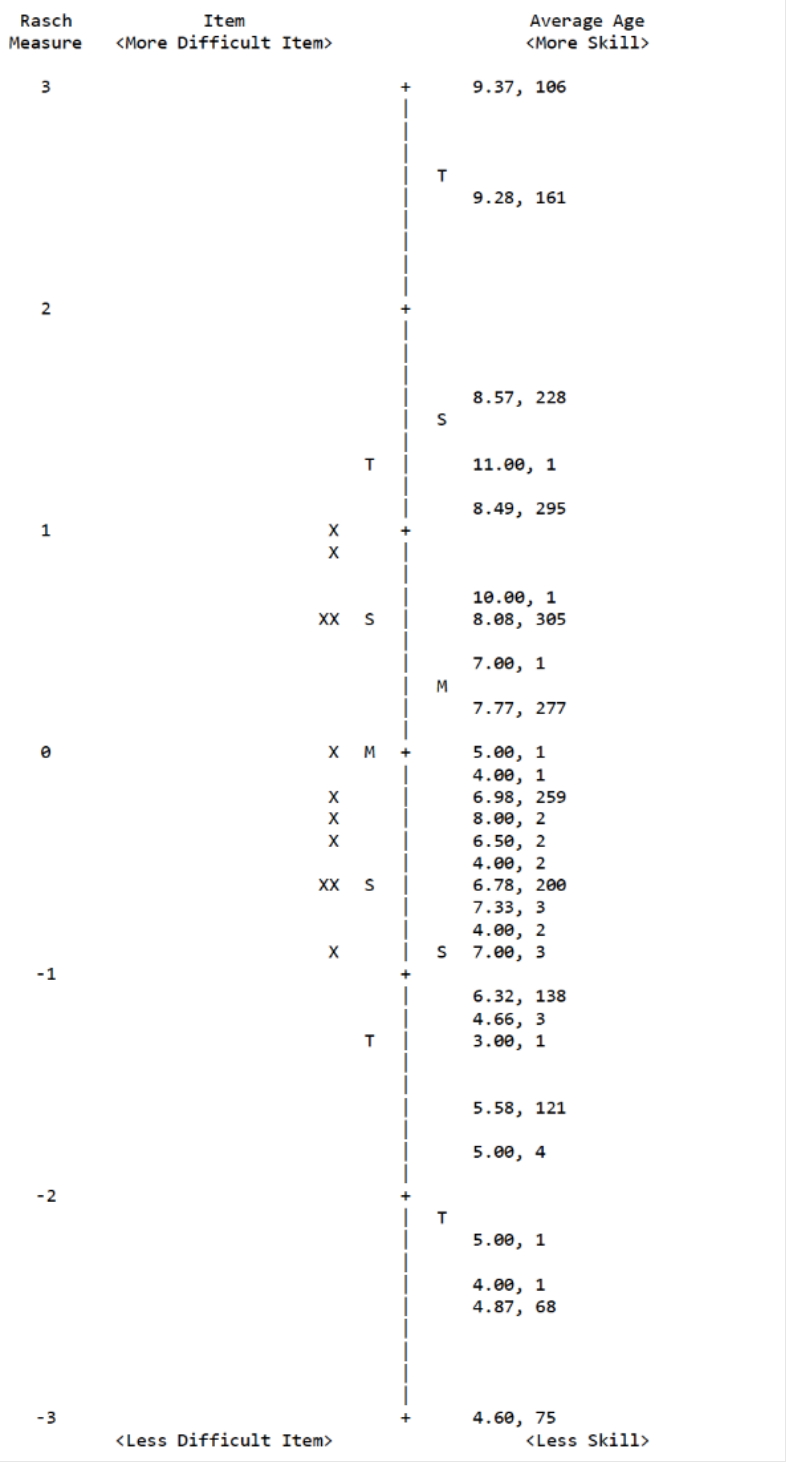



**Figure 4.12**

*Tactile Perception: Oral Person Map*

*Note.* | = latent trait (tactile perception), left side of the figure shows items (X), right side of the figure shows (average age, # of children), M = item/child mean, S = 1 SD, T = 2 SD

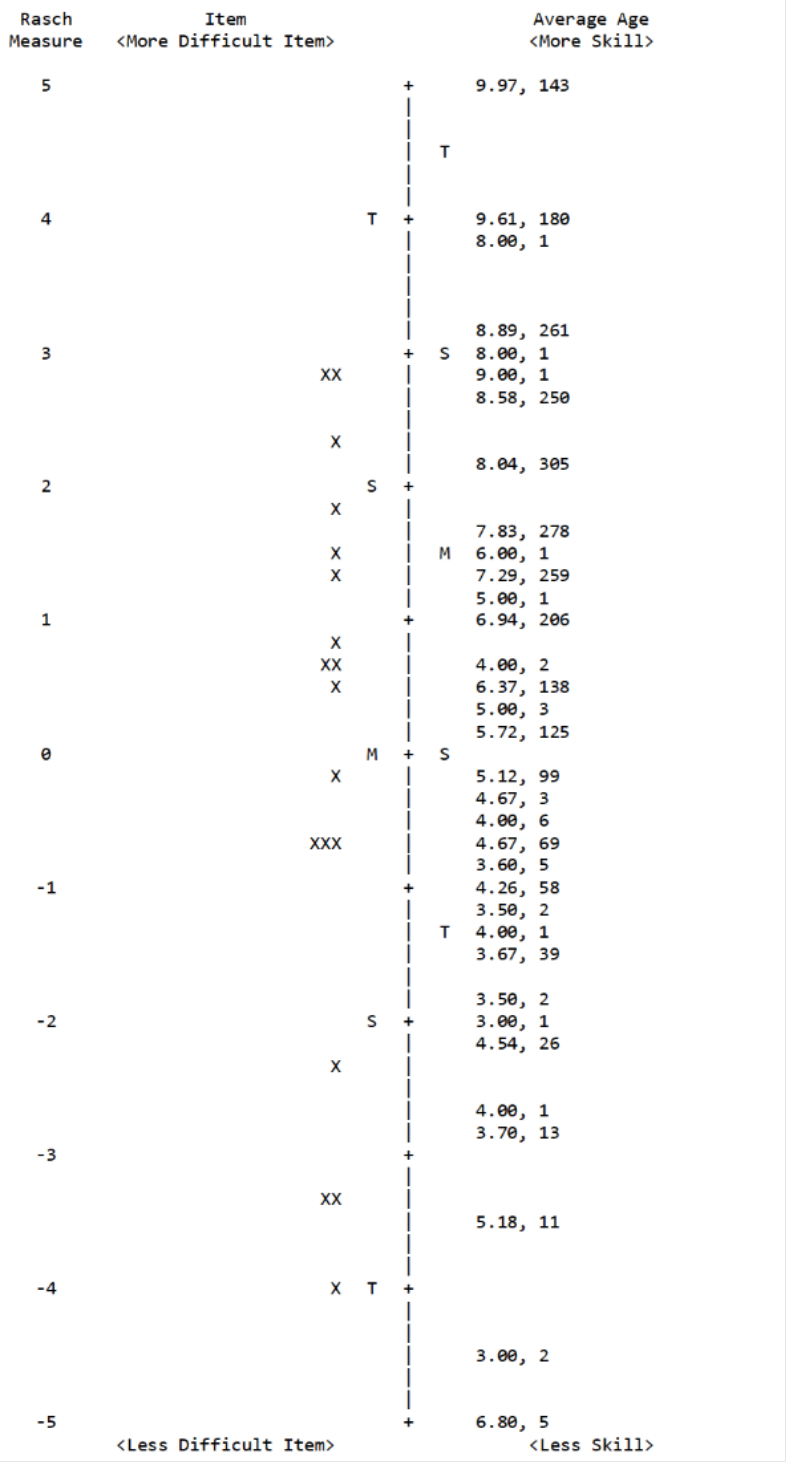



**Figure 4.13**

*Visual Perception: Search Person Map*

*Note.* | = latent trait (visual perception), left side of the figure shows items (X), right side of the figure shows (average age, # of children), M = item/child mean, S = 1 SD, T = 2 SD

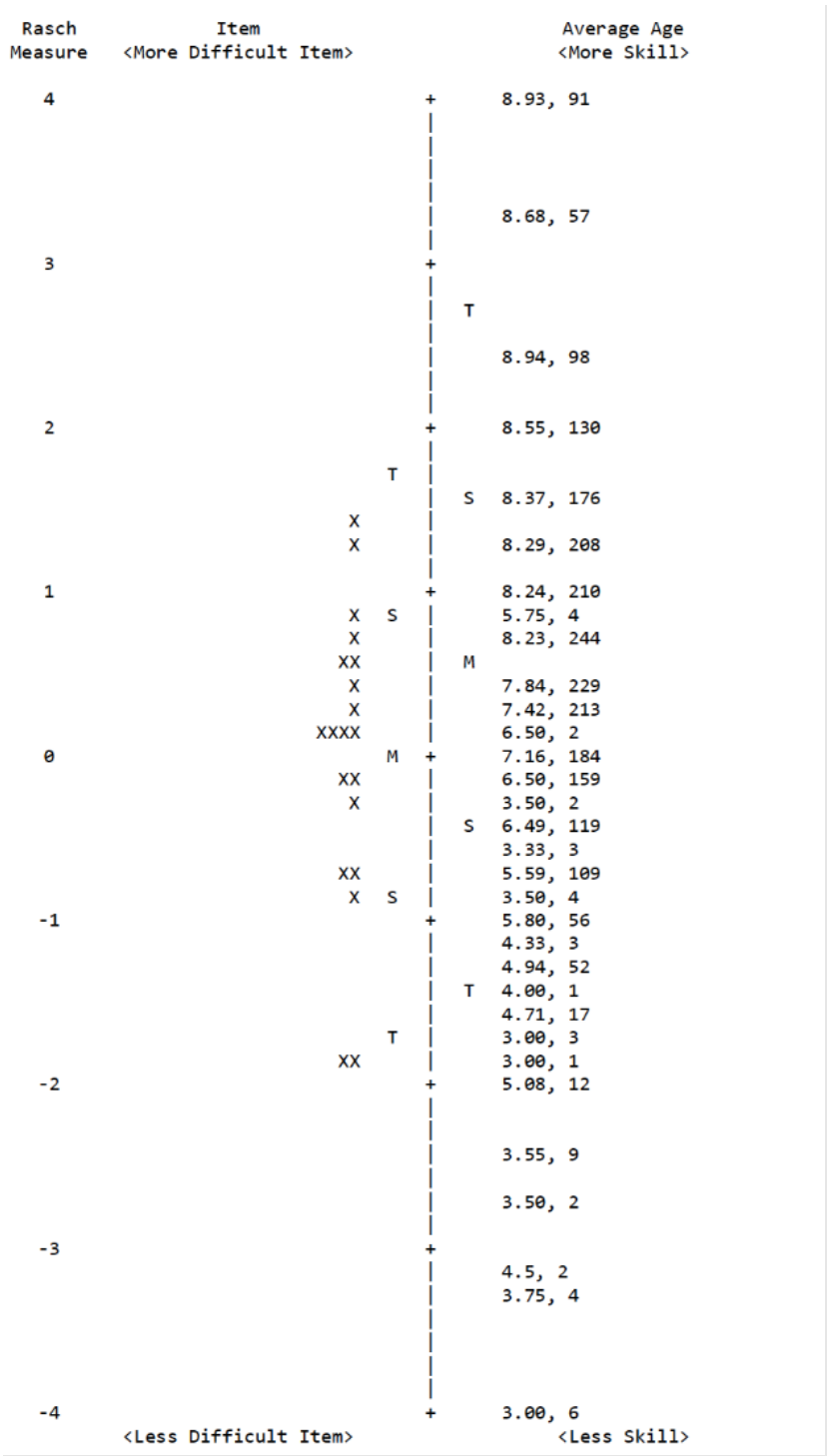



**Figure 4.14**

*Auditory Localization Person Map*

*Note.* | = latent trait (auditory localization), left side of the figure shows items (X), right side of the figure shows (average age, # of children), M = item/child mean, S = 1 SD, T = 2 SD

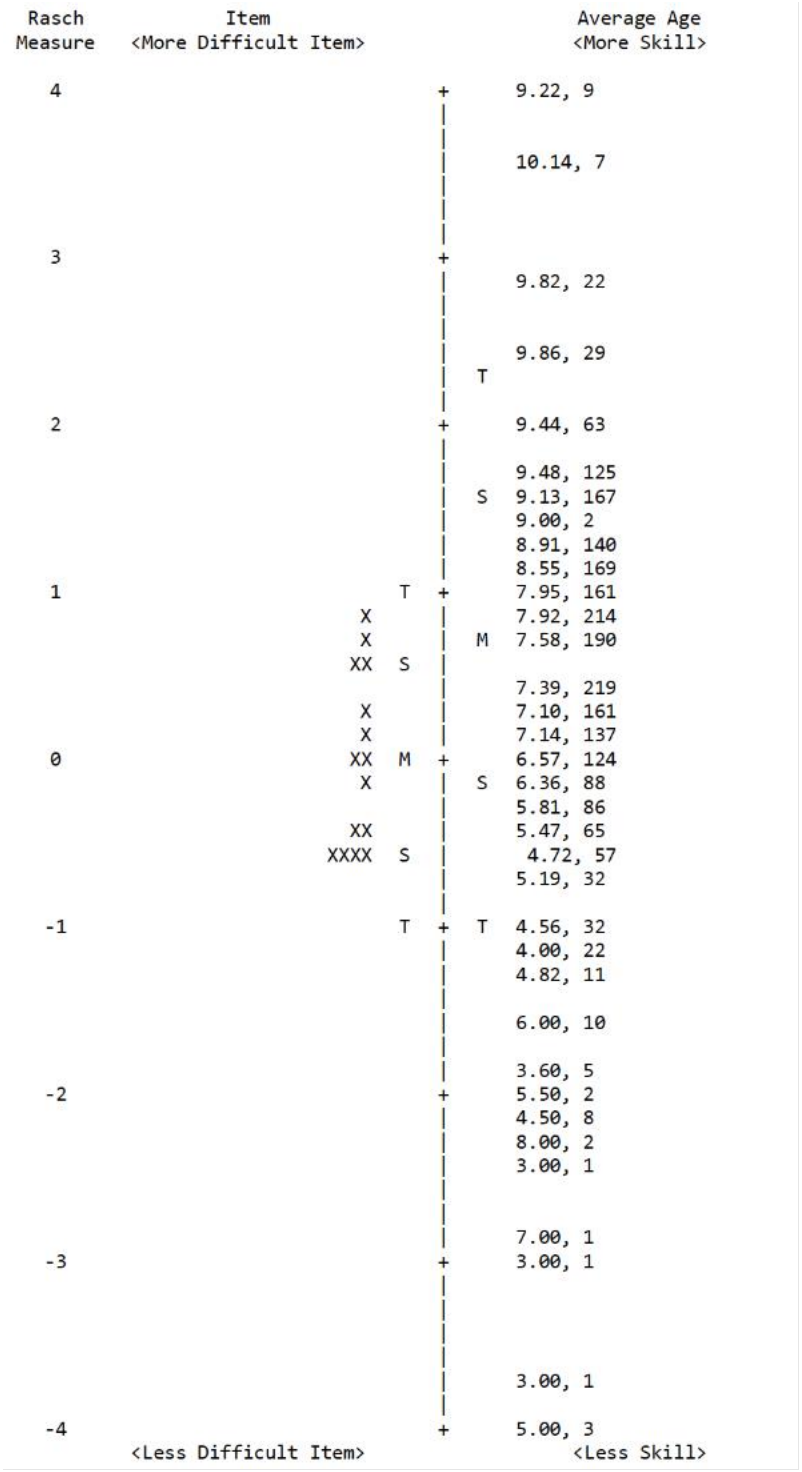



**Figure 4.15**

*Proprioception: Joint Position Person Map*

*Note.* | = latent trait (proprioception), left side of the figure shows items (X), right side of the figure shows (average age, # of children), M = item/child mean, S = 1 SD, T = 2 SD

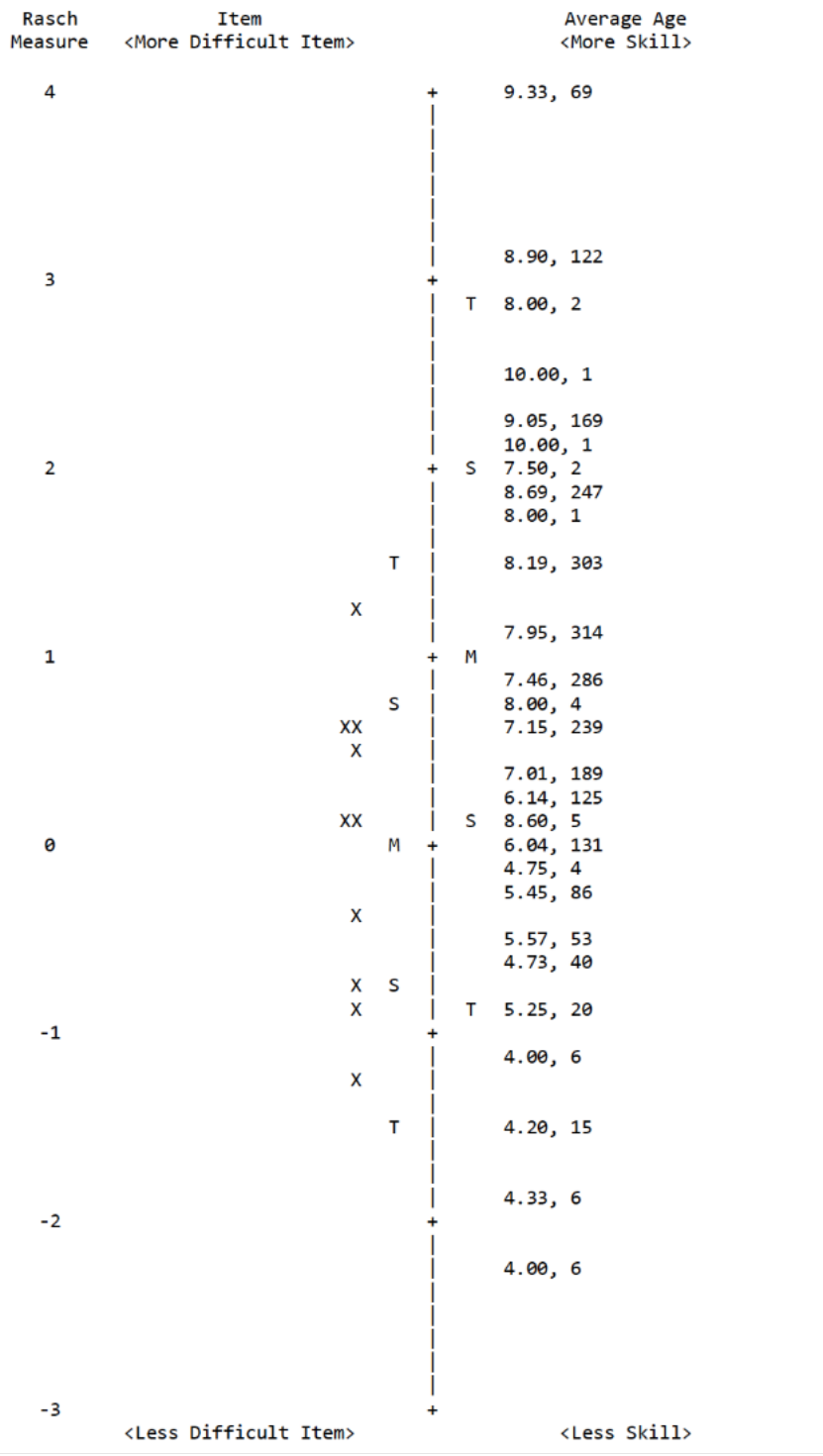



**Figure 4.16**

*Proprioception: Force Person Map*

*Note.* | = latent trait (proprioception), left side of the figure shows items (X), right side of the figure shows (average age, # of children), M = item/child mean, S = 1 SD, T = 2 SD

**Table 4.17**

*Bivariate Correlations Between Age in Months and Sensory Perception Measure Scores*

| Test | Correlation |
|---|---|
| Tactile Perception: Localization | 0.38*** |
| Tactile Perception: Designs | 0.61*** |
| Tactile Perception: Shapes | 0.55*** |
| Tactile Perception: Oral | 0.48*** |
| Visual Perception: Search | 0.59*** |
| Auditory Localization | 0.39*** |
| Proprioception: Joint Position | 0.45*** |
| Proprioception: Force | 0.43*** |

*Note.* *** = p < .001

**Internal Reliability**

Table 4.18 displays reliability indices (person reliability index and strata) for each of the eight sensory perception tests. TPL, TPO and PF fell below our criteria for adequate person reliability index (< .70); only PF demonstrated inadequate strata (< 2.0). All other tests had strong (TPD) or adequate (TPS, VPS, AL, PJP) evidence for internal reliability.

**Table 4.18**

*Rasch Reliability*

| Test | Person Reliability Index | Strata |
|---|---|---|
| Tactile Perception: Localization | 0.63 | 2.07 |
| Tactile Perception: Designs | 0.88 | 4.03 |
| Tactile Perception: Shapes | 0.73 | 2.53 |
| Tactile Perception: Oral | 0.66 | 2.19 |
| Visual Perception: Search | 0.77 | 2.75 |
| Auditory Localization | 0.73 | 2.52 |
| Proprioception: Joint Position | 0.73 | 2.52 |
| Proprioception: Force | 0.59 | 1.93 |

## Discussion

In this study, I used the Rasch model to evaluate evidence for the validity and reliability of normative data collected using eight EASI tests examining aspects of sensory perception. Overall, the study suggests mixed evidence for the psychometric properties of these data.

### Recommended Revisions to the EASI Sensory Perception Tests

Rasch analyses revealed several areas for improvement on the EASI sensory perception tests. The findings reveal the need for significant improvement of two tests: PJP and PF.

#### Proprioception: Joint Position

Rasch analysis of PJP revealed a number of threats to the construct validity and internal reliability of data gathered with this test. Most significantly, although all items fit the model, principal components analysis revealed a significant contrast between items belonging to the Hand and Foot subtests and items belonging to the Two Hand subtests. Large item contrasts suggest that a test does not measure a unified construct (Bond, Yan & Heene, 2020). Multidimensionality may stem from different test demands. On the Hand and Foot subtests, the examiner places the child's limb at an angle and the child must return to the same location. On the Two Hand subtest, the child places both arms within a target on either side of a door and must match the height of their arms. Although the child's vision of the door is occluded, they can still see their upper arms. It is possible that a child with poor perception of joint perception might use visual cues to compensate.

The PJP Wright map revealed that four of the five Two Hand items occupied the same difficulty level along the hierarchy; therefore, they are likely duplicative and add little novel information about the child's proprioceptive perception. The Two Hand subtest requires a complex setup with different materials and room requirements than the other two subtests.

Combining clinical utility and empirical data, I recommend that these items be removed from the PJP overall score.

### Proprioception: Force

The PF test also requires significant revision. Through PCA, I identified a secondary dimension within the data: the Crayon items diverged from the Rolling Bottle items. These two sets of items may measure different aspects of force perception. The Crayon items require the child to set and moderate a degree of force based on a visual stimulus; they receive constant feedback (the lightness of their image). On the Rolling Bottle items, on the other hand, the child must match force applied during two trials; the task comprises a strong praxis component including timing and coordination. Although I theorized that both tasks measured proprioceptive perception of force, I did not find a high correlation between the two contrasts (disattenuated correlation = .27). Possibly the influence of praxis on the Rolling Bottle items outweighs the force perception aspect, causing the two subtests to diverge.

Most current tests designed to measure force perception rely on expensive and highly specific test materials (e.g., Hatzfeld, Kern & Werthschützky, 2010). These tests use everyday materials and allow examiners to understand the child's functional force perception. However, the results of this analysis do not lend confidence to either subtest. At this time, examiners will continue collecting data with these tests, but the EASI authors are not providing normative scores.

### Additional Recommendations

Data from the remaining six tests showed mixed evidence for validity and reliability although most Rasch parameters were acceptable. Data from three items failed to fit on both TPD and VPS, bringing these tests under our threshold of 95% item fit. For TPD, all misfitting

items showed acceptable infit, but poor outfit and very low difficulty; this failure to fit is likely the result of unexpected wrong answers on easy items (Linacre, 2002). Possibly, some children were careless when responding to items that were far below their ability levels. Because all infit parameters were acceptable, I chose to retain these items. For VPS, the sources of failure to fit are less clear. Once again, however, all infit statistics (statistics that attribute more weight to children with ability levels closest to each item's difficulty) met the criteria. Moreover, none of the misfitting items showed notable DIF, nor did they appear to form a secondary dimension on PCA. Therefore, I do not consider these items a major threat to construct validity; I retained them; they should be monitored in future analyses. All remaining tests showed acceptable item fit for 100% of items.

TPD, TPS and VPS fell below our person fit criteria (90%). All three tests approached 90%, with TPD lowest at 84.8%. The children with misfitting data may be impacted by the relative ease of these tests for typically developing children; just a handful of unexpectedly incorrect items could cause a child's data to fail to fit. These mistakes may be caused by factors such as inattentiveness, boredom, or test fatigue rather than a true lack of ability (Smith & Plackner, 2009). Future studies will include both typically developing children and those with sensory integration concerns; I expect better person fit among the children for whom these assessments are designed.

Rating scale analyses, DIF, and PCA supported construct validity for data gathered with TPL, TPD, TPS, TPO, VPS and AL. These results suggest each test measures a unidimensional aspect of sensory perception, and that test items are not significantly impacted by off-dimensional noise (Bond, Yan, & Heene, 2020).

Data from four tests (TPD, TPS, VPS and AL) met our criteria for acceptable reliability; indices for TPL and TPO fell below the accepted threshold (.70). In the Rasch model, both person reliability indices and strata rely highly on sample-item targeting: in other words, when items are too easy or too difficult for the sample population, reliability indices will be lower (Linacre, 2022). This is consistent with the Wright maps for TPL and TPO – many children's ability estimates exceeded even the most difficult items. Although TPS, VPS and AL exceeded the minimum criteria, they did not reach the commonly accepted standard for "good" reliability (.80). As discussed previously, I expect to see improved targeting (and, therefore, higher reliability) among clinical groups and children with known sensory integration concerns.

In addition to sampling other groups, these results suggest a need for more difficult items to reliably measure older children. Gross evaluation of our data suggested that older children tended to score higher than younger children; this is consistent with previous literature in sensory perception (see Bremner & Spence, 2017 and Johnson, 2010 for reviews). It is possible that the EASI items may be too simple to correctly identify older children with deficits. However, I hesitate to recommend additional items given the length of the tests. If the EASI authors choose to create additional more difficult items, I recommend: (1) re-establishing normative data for these items and (2) investigating stop criteria to reduce assessment burden and (3) determining if all subtests should be given to all children.

Additionally, the EASI sensory perception tests may be well-suited to computer adaptive testing (CAT). CAT uses IRT models (such as Rasch analysis) to create parameters for algorithms that select only the questions most relevant to the test-taker's ability level (Linacre, 2000). This approach can significantly reduce assessment burden for children and examiners

(Cella et al., 2007). Previous teams have successfully employed CAT to create item banks and tailored tests for pediatric patients with disabilities (e.g., PEDI-CAT; Haley et al., 2011).

**Developmental Trends Observed in Sensory Perception EASI Scores**

While it is possible that older children (10-12 year olds) with deficits will not be well-measured by the EASI, I am encouraged by the developmental trends observed among the sensory perception tests. For all tests, I observed trends towards increasing scores with increasing age, supporting the construct validity of these tests. Children reached maximal scores on TPL at an average of 8.5 years old, with some children reaching high scores at a much younger age. TPS, TPO, and TPD, however, showed children reaching maximal scores closer to 10 years. This is consistent with theoretical understanding of sensory perception: children must at least be able to localize tactile input before determining more detailed features of tactile stimuli, such as direction of movement (TPD) and texture/shape details (as required for TPS and TPO). Moreover, it is consistent with research using the SIPT (Ayres, 2005), which showed relatively little development between 4 and 8.5 years for Finger Identification and Localization of Tactile Stimuli (SIPT equivalents to TPL) and rapid development on more complex tactile tests such as Manual Form Perception (equivalent to TPS) and Graphesthesia (equivalent to TPD).

I also observed expected developmental trends for VPS, AL, PJP and PFP. VPS age patterns agreed with those observed for Figure Ground (the SIPT equivalent; Ayres, 2005), while PJP agreed with those observed for the SIPT Kinesthesia test. Few researchers have examined the development of auditory localization past infancy, but the authors of a single study conclude that this sensory function increases throughout adolescence (Kühnle et al., 2013). I found no

previous studies examining the developmental trajectory of force perception, but based on other sensory functions, it is encouraging that PFP scores also show a moderate correlation with age.

### Limitations

Although this study used a relatively large, international sample, examiners recruited children based on convenience; many children were known to the examiners in advance of testing. This may have impacted results. Also, most examiners were clinicians working with children who have SI disorders. To prevent a skewed sample, our team asked examiners not to recruit siblings of children in their practice, as siblings may be more likely to share diagnoses and may not represent the typically developing population. However, I recognize that this may have 'over-corrected' the sample and resulted in a sample with higher ability than the true typically developing population.

Our normative data collection efforts took place during the COVID-19 pandemic. As a result of varied restrictions across countries, our sample is not evenly distributed across the included world regions. Data collection is still underway. As the restrictions change, I expect more countries to meet their normative data collection goals, resulting in a more representative population. When sample sizes permit, DIF analyses should be completed by countries and/or world regions to examine whether norms should be stratified by location.

## Conclusions

Preliminary analyses of the normative data collected for the EASI sensory perception tests revealed the need for substantial revisions to two tests: PJP and PF. For the remaining six tests, I found generally good evidence for construct validity and reliability. Most areas of weakness within the tests can be explained by the relatively high ability of our sample population; therefore, our team will examine the psychometric properties of data collected with

children who have clinical SI diagnoses. These results suggest that the normative data form a robust basis for the EASI.

# CHAPTER 5: MANUSCRIPT 3: MOTOR SKILLS

Sensory integration (SI) refers to "the neurological process that organizes sensation from one's own body and from the environment and makes it possible to use the body effectively within the environment" (Ayres, 1972). Sensory integration theory is a scientific theory that postulates the relationships between sensory integration, motor development, and academic/learning disabilities (Bundy & Lane, 2020). SI theory also gives rise to a treatment method – SI therapy – based on the theory that opportunities for enhanced sensation can facilitate adaptive responses (i.e., voluntary, functional movements and responses that allow the child to engage with their environments and participate in meaningful activities).

In sensory integration theory, Ayres (1972) emphasized the importance of childhood motor development as both an indicator and a result of functional sensory integration. Specifically, she emphasized postural-ocular control, bilateral coordination, and balance. Each of these skills follows a relatively predictable progression during early childhood. The ability to maintain static and dynamic postural control and balance against gravity underscores a child's ability to walk, reach, run, and complete many goal-directed tasks and occupations (Ayres, 1972; Keogh & Sugden, 1985; Parham & Cosbey, 2020). Ocular skills are critical to maintaining a stable visual field during dynamic activities (Blanche et al., 2021). Bilateral coordination allows the child to smoothly and continuously use both sides of the body. Development of these skills is driven by opportunities to engage with increasingly demanding environments (e.g., a 6-month-old child rotates a ball in his hands, a preschooler throws the ball in the air and catches it, and a 7-year-old runs down a soccer field, kicking the ball).

For some children, however, these basic motor skills fail to follow a normal developmental progression. In SI theory, we conceptualize these skills as intimately linked to perception and integration of sensory information (Ayres, 1972). Postural-ocular control, for example, relies on the child's ability to sense the position and movement of the body in space (i.e., vestibular and proprioceptive perception). Failure to accurately perceive and/or respond to sensory information may result in poor motor skills.

Assessment of motor skills, therefore, provides critical insight into sensory processing. Bundy and Lane (2020) described two conditions that fall under the scope of SI deficits: Vestibular, Bilateral Integration and Sequencing (VBIS) and Somatodyspraxia. They hypothesized that VBIS deficits are primarily driven by poor central processing of vestibular and proprioceptive sensations (i.e., poor postural-ocular control). While motor control in static or simple positions/movements may be sufficient for completing basic tasks, more complex activities (such as those requiring smooth coordination of two arms or legs) may challenge the child.

Somatodyspraxia, characterized by severe deficits in planning and executing motor tasks, is likely related to deficits across the somatic senses (tactile, vestibular and proprioceptive). These senses are responsible for the child's internal body scheme – without proper discrimination in the somatic senses, the child does not develop normal motor and downstream praxis abilities. Even very simple tasks may be difficult for the somatodyspraxic child. While both conditions are linked to SI deficits, they have different functional implications and require nuanced treatment approaches. Therefore, therapists using SI treatment approaches must carefully evaluate children's motor development to plan useful intervention.

## Clinical Observations

Previously, therapists using SI therapy relied primarily upon Ayres's recommended clinical observations to assess motor skills (Blanche, Reinoso, & Kiefer, 2020). Tasks such as supine flexion, prone extension and reaching allow the therapist to observe the child's balance and postural control under dynamic and challenging circumstances. Ocular tasks, such as convergence, divergence, localization and tracking, provide insight into vestibular-ocular integration. Jumping jacks, stride jumps, and symmetrical arm/hand movements provide insight into bilateral integration. Slow ramp movements, where the child follows the examiners' motions closely, demonstrate the child's ability to move fluidly and with control.

While Ayres's clinical observations have remained relatively consistent over the years, they are neither standardized nor normed against typically developing children. Several attempts have been made to standardize the observations. Examples of these tests include Clinical Observations of Motor and Postural Skills (COMPS; Wilson et al., 1999), Sensory Integration Clinical Observations (SICO; May-Benson & Teasdale, 2021), and Structured Observations of Sensory Integration – Motor (SOSI-M; Blanche, Reinoso & Kiefer, 2021). Despite these attempts, all fall somewhat short of clinicians' needs when assessing children with potential sensory integration deficits. COMPS omits ocular items, while both SICO and SOSI-M diverge into observations of praxis, which I argue should be evaluated and scored separately from motor abilities. Furthermore, standardization groups for SICO, SOSI-M and COMPS are limited to North American populations (see Literature Reiew for details).

## The Evaluation in Ayres Sensory Integration (EASI)

To fulfill the assessment needs of clinicians working with children who have suspected sensory integration deficits, our team developed the Evaluation in Ayres Sensory Integration

(EASI). The EASI is a norm-referenced, standardized suite of tests that evaluate SI functions including motor skills, praxis, sensory perception, and sensory reactivity. Four of these tests are relevant to the evaluation of motor and postural skills. Postural Control (PC), Balance (BAL), Ocular Motor (OM) and Bilateral Integration (BI). Table 5.1 describes these tests in more detail.

**Table 5.1**

EASI Motor Tests

| Test | Description | Scoring | # of Items |
|---|---|---|---|
| Postural Control (PC) | Child assumes and maintains a variety of positions and completes reaching tasks | **Accuracy Items [1] (all except Items 7 and 13)**<br>2: Child demonstrates good postural control to complete the task/maintain the position<br>1: Child demonstrates adequate postural control with some awkwardness, hesitation or instability<br>0: Child does not demonstrate adequate postural control<br>**Time Items (Items 7 and 13)**<br>2: Child maintains position for 30 seconds<br>1: Child maintains position for 10-29 seconds<br>0: Child maintains position for less than 10 seconds | 19 |
| Balance (BAL) | Child assumes and maintain positions with eyes open and closed (e.g., standing on one foot) | **Items 1 and 2**<br>1: Child maintains position for 10 seconds or more<br>0: Child maintains position for less than 10 seconds<br>**Items 3-12**<br>2: Child maintains position for 10 seconds or more<br>1: Child maintains position for 5 to 9 seconds<br>0: Child maintains position for less than 5 seconds | 12 |
| Ocular Motor (OM) | Child demonstrates ocular skills such as ocular pursuits, stabilization and quick localization | 2: Child's eyes/head move smoothly and precisely[1]<br>1: Child's eyes/head move functionally, but there is some hesitation or jerkiness<br>0: Child's eyes/head do not move in a way that completes the task | 14 |
| Bilateral Integration (BI) | Child performs sequences of activities that use both sides of the body, as demonstrated by the examiner | 2: All actions are synchronized between both sides of the body correctly, smoothly, and rhythmically.<br>1: All actions are correct, but slightly jerky or dysrhythmic.<br>0: Actions are not synchronized, incorrect, or are jerky or dysrhythmic. | 15 |

[1]Definitions of categories 2, 1 and 0 are slightly different for each item; however, the descriptors here provide an overview of the functional levels intended with each category.

I grouped these four tests into a single manuscript because the primary constructs they measure are most closely related to motor ability. It is worth noting that some tests have significant overlap with praxis, another construct measured by the EASI. This is logical, as praxis is the process of enacting motor skills appropriately to accomplish a target result or behavior (Cermak & May-Benson, 2020). In these tests, the EASI developers aimed to primarily isolate motor skills. However, praxis is likely to impact test scores. Testers are encouraged to eliminate praxis demands as much as possible on these tests; they can use verbal and physical cues to ensure that the child understands the task.

**The Present Study**

In this study, I conducted Rasch analyses of each of the four EASI motor tests. Through these analyses, I explored the evidence for construct validity and internal reliability of normative data collected using these tests. Establishing evidence for validity and reliability is a critical step in bringing a novel assessment into clinical practice. In order to adopt a new test, clinicians must be confident that the items produce data that accurately and consistently represents the strengths and limitations of the children they serve.

Rasch analysis is a model in the item response theory (IRT) family; this model allows us to examine individual items and each test as a whole (Bond, Yan & Heene, 2020). Using Rasch analysis, I can establish that test items form a unidimensional construct – a critical step in establishing construct validity. Furthermore, I can evaluate the usefulness of these items for evaluating the children for whom this test is designed, and the replicability (i.e., reliability) of data collected. Specifically, I examined the following research questions:

(1) What is the evidence for construct validity of the data collected using each of the four EASI motor tests?

a. Do the test items demonstrate uniformly positive point-measure correlations (i.e., do scores on each item correlate with overall test score?)
b. Do 95% of items demonstrate adequate fit to the Rasch model?
c. Do 90% of children demonstrate adequate fit to the Rasch model?
d. Do the Rasch-generated step thresholds within rating scales progress in an orderly fashion?
e. Does a Rasch principal components analysis of standardized residuals reveal meaningful secondary dimensions in the data?
f. Does a differential item functioning analysis reveal invariance in item difficulty for male and female children?
g. Do the items form a logical hierarchy with sufficient item difficulty variation to match sample ability levels?
h. Do test-takers form a logical developmental hierarchy (i.e., do scores increase with increasing age?)

(2) What is the evidence for internal reliability of data collected using the 4 EASI motor tests?

a. Do the data demonstrate adequate internal reliability, based on the Rasch person reliability index?
b. Do the data reliably distinguish at least two levels of motor skill, based on the number of strata associated with the measure?

## Methods

I used Rasch analysis to evaluate evidence for validity and reliability of international normative data collected using the four EASI motor tests. I conducted a separate analysis for each of the tests; in this report, I summarize the results of these analyses.

### Participants

I drew data from the EASI International Normative Data Collection Project (maintained by the Collaboration for Leadership in Ayres Sensory Integration [CLASI]). The dataset comprises 2563 children between the ages of 3-12 years. Inclusion criteria for normative data collection included: (1) chronological age between 3 years and 0 months – 12 years and 11 months; (2) typical development; (3) no known medical, educational, mental health, or other developmental concerns. Exclusion criteria were: (1) known medical, educational, mental health, or other developmental concerns; (2) identified as having sensory integration concerns by OT, PT, or SLP; (3) receive(s/d) therapy services for learning disorders, ASD, ADHD, speech/language delays, regulatory issues, hypotonia, or DCD; (4) siblings who meet any of these exclusion criteria. Not all children completed every test; individual test sample sizes are reflected in Table 5.6. In Appendix C, I described the entire sample. The Rasch model is robust against missing data; therefore, I included children who did not complete all items within tests. I only omitted children with less than 50% of items completed.

### Procedure

All normative data collectors (examiners) completed an 8- to 10-hour online training course that covered EASI testing and scoring. At the conclusion of training, they completed a series of online scoring quizzes, achieving at least 80% accuracy against scores completed by a gold standard observer. examiners conducted EASI tests in locations convenient for children and

families. This included clinics, children's homes, and research laboratories. Most examiners gave all 21 tests during a single 3- to 4-hour session; however, some required multiple sessions. Most common reasons for multiple sessions included scheduling conflicts or children's limited tolerance for extended testing. examiners uploaded all data into a secure RedCap database managed by the Collaboration for Leadership in Ayres Sensory Integration (CLASI).

## Data Analysis

The Rasch model is a latent trait psychometric model that converts ordinal-level data (e.g., EASI) to interval-level measures (Bond, Yan & Heene, 2020). Both person ability and item difficulty are estimated along the same log-odds unit ("logit") scale. Rasch is based upon two complementary assumptions; these assumptions can be expressed in terms of the EASI sensory perception tests, that: (1) easier items (i.e., items that require less developed motor skills) were easier for all children, and (2) children with more developed motor skills could complete harder items.

### Model Selection

The Rasch model includes several sub-models for analyzing data with varied rating scales (Bond, Yan & Heene, 2020). While the original model applied only to dichotomous data, the rating scale models (RSMs) are appropriate for data with ordinal rating scales; these models provide fit statistics and logit calibrations for rating scale categories in addition to persons and items. This model can be further divided into the standard RSM (all items share a polytomous rating scale) and the grouped RSM (items share multiple polytomous rating scales). Following Linacre's (2000) guidance, I selected RSM for OM and BI and grouped RSM for both PC and BAL.

### Construct Validity

In addition to generating item, person and rating scale calibrations, Winsteps allows users to evaluate several indicators that suggest the extent to which the data fit the Rasch model. I examined these indicators for evidence of construct validity: point-measure correlations, goodness-of-fit statistics, rating scale thresholds, item hierarchies, person hierarchies, PCA, and DIF statistics.

***Point-measure Correlations.*** To determine if each item corresponds with the latent variable (i.e., that a higher score corresponds to improved motor skills), I examined Pearson point-measure correlation coefficients between observations and item measure. In the Rasch model, positive point-measure correlations suggest that items align with the construct (Bond, Yan & Heene, 2020). Of note, the magnitude of these correlations is less important than the directionality. To establish construct validity, all point-measure correlations should be positive.

***Goodness-of-fit Statistics (Items).*** I examined two kinds of mean-square goodness-of-fit-statistics generated by Winsteps: infit and outfit. Infit statistics are "inlier-sensitive" or information-weighted to reduce the influence of off-target responses (i.e., children whose overall scores are far from the item measure). Outfit statistics are unweighted and typically reflect fit problems due to outliers. Mean-squares show the amount of distortion of the measurement system. Ideal mean-square value is 1.0. Values between 0.7 and 1.3 suggest adequate fit to the Rasch model (Linacre, 2002).

To demonstrate sufficient evidence for construct validity, at least 95% of items on each test should show adequate fit to the Rasch model. For tests with fewer than 20 items, a single misfitting item would fall below this threshold. Because I might expect at least one item to fail to

fit due to chance alone, I expected *either* 95% of items to fit the model, or *no more than one* item to misfit.

***Goodness-of-fit Statistics (Children).*** Person fit statistics are calculated and interpreted in the same way as item fit statistics. As a rule, people often behave less predictably than items (Bond, Yan, & Heene, 2020). Further, given that I had relatively few items compared to people, I selected less stringent criteria for acceptable mean-squares (0.5 to 1.5) compared to the criteria for items. People who overfit the model (i.e., behave too predictably, mean-square < 0.5) are unlikely to distort or degrade the measurement system (Linacre, 2015). Therefore, I only considered children misfitting if they underfit the model (MnSq > 1.5). I expected that 90% of children would fit the model for each test for evidence of strong construct validity.

***Rating Scale Analysis.*** For all rating scales, I examined (1) rating scale goodness-of-fit statistics. Mean-square between 1.3 - 0.7 suggested fit to the Rasch model; Linacre, 2021). (2) Observed average person measure associated with each category. The observed average person measure associated with each category should demonstrate orderly progression; the lowest category should correspond with the lowest average person measure (Bond, Yan & Heene, 2020). (3) Andrich thresholds (i.e., the person ability measure at which a person is equally likely to use two adjacent categories). Andrich thresholds should progress in an orderly fashion, such that the lowest step threshold corresponds to the threshold between the two lowest categories and so forth (Bond, Yan & Heene, 2020; Linacre, 2018).

***Principal Components Analysis.*** While goodness-of-fit statistics and other evidence described above examine the extent to which the construct is unidimensional, PCA provides evidence of the strength of additional dimensions in the data (i.e., multidimensionality). PCA deconstructs model residuals to identify additional dimensions in the data (i.e., item response

patterns not explained by the Rasch model). Eigenvalues estimate the strength of these dimensions (called contrasts). I considered contrasts to be strong enough to refute unidimensionality of the construct if the following conditions were met (Linacre, 2018): (1) There were contrasts with eigenvalues > 2 (i.e., with the strength of more than 2 items); and (2) Item subsets within contrasts demonstrate disattenuated correlations < 0.57, indicating that item subsets likely measure different latent variables. According to Wright and Stone (1979), unidimensionality is essential to good construct measurement; evidence of multiple dimensions suggests that the items should be scored as multiple, separate instruments.

***Measurement Invariance.*** I used Rasch DIF analyses to examine the measurement invariance of the motor tests based on sex (i.e., that test items are not biased based on the child's sex). Using the Rasch-Welch DIF method (Linacre, 2020), I compared item difficulty estimates for males and females. DIF contrasts (i.e., the difference between difficulty estimates) for males and females should be no larger than .43 logits (Zwick, Thayer and Lewis, 1999) to be considered negligible. I also conducted *t*-tests of item difficulty to examine the likelihood that DIF could be caused by chance alone. Items with both DIF contrast > .43 *and* $p < .05$ should be considered problematic and should be removed or targeted for revision. Given the large sample size used in this study, I did not consider items to show bias if contrasts were significant but smaller than .43 logits.

***Item Hierarchy.*** I assessed item hierarchies in two ways. First, I compared the mean item measure and the mean person measure. In the Rasch model, the mean item measure is set at 0.0 logits. Mean person measure close to 0.0 indicates a match between the sample's sensory integration ability and the scale difficulty (Bond, Yan & Heene, 2020). Second, I visually inspected the Winsteps-generated Wright maps. The Wright map provides a hierarchy of items

and persons along a logit scale, ranging from lowest to higher measures. These items should be ordered logically so that theoretically-more-difficult items are associated with higher item difficulty measures. Therefore, I descriptively examined the item hierarchies to ensure that they matched theoretical expectations. I also examined the Wright maps to evaluate the spread of items; large gaps in item difficulty indicate a need for more items, while items grouped together suggest redundancy.

***Person Hierarchy.*** I assessed person hierarchies in two ways. The Winsteps-generated person maps show child scores along a continuous, interval-level scale on the right side of the figure and items along the left side (see, for example, Figure 5.1). The interval scale is broken into .2 logit levels. I averaged the ages of children on each level and visually inspected the map for evidence that average age increases with increasing scores. Second, I evaluated the strength of this relationship by conducting bivariate Pearson correlations between Rasch-generated child measure scores and children's age in months. Given the developmental nature of sensory integration constructs, I expected at least moderate correlation coefficients ($\geq$.30; Cohen [1988]). I confirmed normality of all variables (age in months and EASI measure scores) using the methods described by Kim (2013) for large sample sizes ($n > 300$).

**Internal Reliability**

I evaluated internal reliability based on two Winsteps-generated indices. The first, person reliability index, is the Rasch equivalent to Cronbach's alpha and represents the amount of variance that can be reproduced by the Rasch model (Wright & Masters, 1982). Person reliability index greater than 0.80 suggests strong evidence for internal reliability; greater than .70 is adequate (Bond, Yan & Heene, 2020).

The strata value is an additional measure of reliability that represents the number of levels of ability that the measure can distinguish (Wright & Masters, 1982). Winsteps generates a separation index (*G*), which I converted to strata using the formula:

$$Strata = \frac{4G + 1}{3}$$

Strata should be at least 2.0 to establish evidence for sufficient internal reliability (Bond, Yan & Heene, 2020). Given the developmental nature of sensory integration and the large age range of our sample, I expect higher strata values (i.e., I expect more levels of ability to be represented by the items). Therefore, I will consider strata values acceptable at 2.0 and strong at 3.0 or more.

## Results

### Construct Validity

Tables 5.2-5.5 display item measures, fit statistics, and DIF statistics. All items showed positive point measure correlations. Data from one PC item failed to fit the model (95.2% of items fit). Data from three items on BAL failed to fit the model (75% of items fit). For OM and BI, all items fit.

**Table 5.2**

*Postural Control Item Measures, Fit Statistics and DIF Statistics*

| Item Description | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
| PE Assuming | PC 1 | -0.03 (0.05) | 0.87 | 0.73 | 0.62 | 0.02 |
| PE Head | PC 2 | -0.23 (0.05) | 0.97 | 1.10 | 0.51 | 0.07 |
| PE Upper trunk | PC 3 | 0.03 (0.05) | 0.87 | 0.83 | 0.59 | -0.09 |
| PE Thighs | PC 4 | 0.93 (0.04) | 0.96 | 0.90 | 0.67 | 0.10 |
| PE Knees | PC 5 | 0.36 (0.05) | 0.92 | 0.83 | 0.63 | 0.00 |
| PE Maintaining | PC 6 | 0.41 (0.04) | 0.82 | 0.74 | 0.66 | -0.11 |
| PE Time | PC 7 | 1.68 (0.04) | 1.03 | 1.02 | 0.67 | -0.15 |
| SF Assuming | PC 8 | 0.26 (0.05) | 0.90 | 1.04 | 0.58 | 0.00 |
| SF Head | PC 9 | 0.04 (0.05) | 0.91 | 0.90 | 0.57 | 0.15 |
| SF Upper trunk | PC 10 | 0 (0.05) | 0.93 | 0.91 | 0.57 | -0.14 |

| Item Description | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
| SF Legs | PC 11 | -0.68 (0.06) | 1.10 | 1.26 | 0.43 | 0.13 |
| SF Maintaining | PC 12 | 0.16 (0.05) | 0.79 | 0.73 | 0.63 | 0.00 |
| SF Time | PC 13 | 1.66 (0.04) | 1.01 | 0.98 | 0.69 | 0.00 |
| Head lag | PC 14 | -1.61 (0.07) | 1.25 | 1.14 | 0.34 | 0.00 |
| Ball | PC 15 | -0.54 (0.05) | 1.16 | **1.39** | 0.43 | 0.00 |
| Robot arms | PC 16 | 1.28 (0.04) | 1.22 | 1.28 | 0.58 | 0.05 |
| Reaching While Standing (Right) | PC 17 | -1.55 (0.07) | 1.16 | 1.09 | 0.37 | 0.17 |
| Reaching While Standing (Left) | PC 18 | -1.31 (0.07) | 1.21 | 1.30 | 0.38 | -0.12 |
| Reaching While Kneel-Standing (Right) | PC 19 | -0.85 (0.06) | 1.17 | 1.28 | 0.41 | -0.05 |

Notes: PE = Prone Extension; SF = Supine Flexion

**Table 5.3**

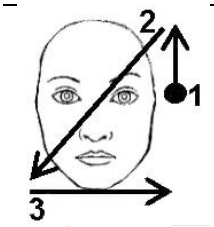
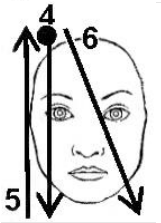
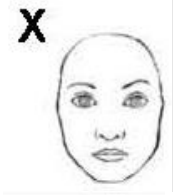
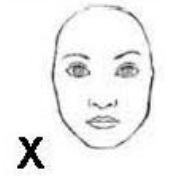
*Balance Item Measures, Fit Statistics and DIF Statistics*

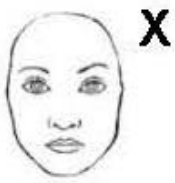
| Item Description | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
| Stand with feet together, toes even, eyes open | BAL 1 | -4.56 (0.17) | 1.22 | **9.90** | 0.23 | 0.13 |
| Stand with feet together, toes even, eyes closed | BAL 2 | -3.15 (0.11) | 1.25 | **9.90** | 0.29 | -0.02 |
| Stand on right foot, eyes open | BAL 3 | -1.2 (0.05) | 0.86 | 0.71 | 0.68 | 0.15 |
| Stand on left foot, eyes open | BAL 4 | -1 (0.05) | 0.91 | 0.84 | 0.68 | 0.00 |
| Stand on right foot, eyes closed | BAL 5 | 1.35 (0.04) | 0.89 | 0.98 | 0.74 | 0.02 |
| Stand on left foot, eyes closed | BAL 6 | 1.41 (0.04) | 0.91 | 0.88 | 0.73 | 0.00 |
| Heel to toe, eyes open | BAL 7 | -1.71 (0.05) | 0.98 | 1.09 | 0.62 | 0.00 |
| Heel to toe, eyes closed | BAL 8 | 0.07 (0.04) | 1.12 | 1.22 | 0.68 | -0.10 |
| Stand on toes, eyes open | BAL 9 | -0.41 (0.04) | 0.98 | 0.79 | 0.70 | 0.00 |
| Stand on toes, eyes closed | BAL 10 | 1.25 (0.04) | 1.02 | 0.97 | 0.72 | -0.15* |
| Stand on toes on one foot, eyes open | BAL 11 | 3.14 (0.04) | 0.90 | 1.01 | 0.68 | 0.00 |

| Item Description | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
| Stand on toes on one foot, eyes closed | BAL 12 | 4.81 (0.06) | 1.15 | **2.66** | 0.51 | -0.23 |

**Table 5.4**

*Ocular Motor Item Measures, Fit Statistics and DIF Statistics*

| Item Description | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
|  | OMP 1 | -0.57 (0.06) | 0.90 | 0.75 | 0.57 | 0.10 |
| | OMP 2 | 0.42 (0.05) | 0.88 | 0.94 | 0.66 | -0.09 |
| | OMP 3 | 0.07 (0.05) | 0.95 | 0.88 | 0.62 | 0.18 |
|  | OMP 4 | -0.3 (0.06) | 0.91 | 0.82 | 0.59 | 0.10 |
| | OMP 5 | 0.04 (0.05) | 0.87 | 0.85 | 0.63 | 0.09 |
| | OMP 6 | 0.91 (0.04) | 0.95 | 0.99 | 0.68 | 0.00 |
| Vertical Plane | OMS 1 | 0.96 (0.04) | 1.15 | 1.10 | 0.66 | 0.00 |
| Horizontal Plane | OMS 2 | 0.96 (0.04) | 1.26 | 1.19 | 0.64 | 0.12 |
| Convergence | OMS 3 | 0.02 (0.05) | 1.15 | 1.12 | 0.55 | -0.27* |
| Divergence | OMS 4 | 0.29 (0.05) | 1.06 | 1.04 | 0.60 | 0.00 |
|  | OMQL 1 | -0.85 (0.07) | 0.97 | 1.18 | 0.49 | 0.00 |
|  | OMQL 2 | -0.82 (0.07) | 0.92 | 0.98 | 0.51 | -0.10 |

| Item Description | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
|  | OMQL 3 | -0.38 (0.06) | 1.00 | 1.06 | 0.53 | -0.14 |
|  | OMQL 4 | -0.74 (0.07) | 0.94 | 0.86 | 0.52 | -0.06 |

OMP = Ocular Motor Pursuits; OMS = Ocular Motor Stabilization; OMQL = Ocular Motor Quick Localization

**Table 5.5**

*Bilateral Item Measures, Fit Statistics and DIF Statistics*

| Item Description | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
| Tap hands on thighs, R, L, R, L, R, L | BI 1 | -0.82 (0.05) | 1.12 | 1.03 | 0.52 | -0.08 |
| Tap hands on thighs, L-L, R-R, L-L, R-R | BI 2 | 0.08 (0.04) | 0.98 | 0.91 | 0.64 | 0.05 |
| Tap hands on thighs, L-R, R, R-L, L | BI 3 | 1.88 (0.03) | 1.21 | 1.20 | 0.65 | -0.21* |
| Pronate then supinate hands/forearms, 3 times | BI 4 | -0.57 (0.04) | 1.04 | 1.08 | 0.54 | 0 |
| Clap, R hand to L shoulder, Clap, L hand to R shoulder | BI 5 | -0.08 (0.04) | 1.06 | 0.97 | 0.60 | 0 |
| Arms out to side, then bring fingers to nose simultaneously, 2 times | BI 6 | -0.73 (0.05) | 0.90 | 1.03 | 0.54 | 0.06 |
| Horizontal circles in air, palms down, 3 times | BI 7 | -0.28 (0.04) | 0.94 | 0.98 | 0.58 | 0.1 |
| Vertical circles in air, palms forward, 3 times | BI 8 | -0.21 (0.04) | 0.90 | 0.88 | 0.60 | 0.16* |
| Bilateral sequential finger touching | BI 9 | 0.97 (0.03) | 0.99 | 0.93 | 0.68 | 0.1 |
| Marching R, L, R, L, R, L | BI 10 | -1.18 (0.05) | 1.14 | 1.08 | 0.47 | 0 |
| Side step R over L, 3 times | BI 11 | -0.16 (0.04) | 1.13 | 1.18 | 0.56 | -0.14 |
| Jumping jacks, 4 times | BI 12 | 0.51 (0.04) | 0.99 | 1.00 | 0.64 | 0 |
| Swing arms up and down from sides, then out to sides, 2 times | BI 13 | 0.53 (0.03) | 0.95 | 0.94 | 0.65 | -0.09 |

| Item Description | Item | Measure (Standard Error) | Infit MnSq | Outfit MnSq | Point Measure Correlation | DIF Contrast |
|---|---|---|---|---|---|---|
| Lift L leg up, then R leg, 2 times | BI 14 | 0.28 (0.04) | 0.97 | 0.97 | 0.63 | 0.12 |
| Lift R leg to R side, then L leg to L side, 2 times | BI 15 | -0.2 (0.04) | 0.84 | 0.87 | 0.61 | 0.04 |

All tests fell below our 90% threshold for person fit (Table 5.6). However, at least 82% of children fit the model for all four tests.

**Table 5.6**

*Person Fit Analysis*

| Test | Children with Misfitting Data | Total Number of Children | % fitting children |
|---|---|---|---|
| Postural Control | 341 | 2463 | 86.2% |
| Balance | 426 | 2459 | 82.7% |
| Ocular Motor | 258 | 2193 | 89.5% |
| Bilateral Integration | 369 | 2065 | 84.8% |

Table 5.7 displays the results of rating scale analyses. All scale categories showed acceptable fit statistics, Andrich thresholds proceeded appropriately, and category thresholds advanced as expected.

**Table 5.7**

Rating Scale Analysis

| Item Type | Rating Scale Category | % Used | Infit MnSq | Outfit MnSq | Andrich Threshold[1] | Observed Average |
|---|---|---|---|---|---|---|
| Postural Control | | | | | | |
| Accuracy | 0 | 3.89% | 0.96 | 1.08 | – | -0.20 |
| | 1 | 20.50% | 0.99 | 1.02 | -1.07 | 1.36 |
| | 2 | 75.61% | 1.00 | 1.00 | 1.07 | 2.99 |
| Maintaining Position | 0 | 16.28% | 0.94 | 0.90 | – | -0.91 |
| | 1 | 34.87% | 1.05 | 1.02 | -0.98 | 0.46 |
| | 2 | 48.85% | 1.08 | 1.07 | 0.98 | 1.43 |
| Balance | | | | | | |
| | 0 | 4.21% | 1.25 | 9.90 | – | 3.15 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Maintaining Position (Dichotomous) | 1 | 95.79% | 1.21 | 1.12 | – | 5.21 |
| Maintaining Position (Trichotomous) | 0 | 31.82% | 1.00 | 1.12 | – | -2.33 |
| | 1 | 18.41% | 0.90 | 0.87 | -0.40 | -0.18 |
| | 2 | 49.77% | 0.98 | 1.67 | 0.40 | 2.58 |
| Ocular Motor | | | | | | |
| Accuracy | 0 | 4.84% | 1.00 | .98 | – | -.22 |
| | 1 | 13.11% | .99 | .96 | -.56 | 1.16 |
| | 2 | 82.06% | 1.04 | 1.04 | .56 | 2.37 |
| Bilateral Integration | | | | | | |
| Accuracy | 0 | 14.99% | 1.03 | 1.07 | – | -.56 |
| | 1 | 18.29% | .96 | .94 | -.17 | .65 |
| | 2 | 66.72% | 1.01 | 1.01 | .17 | 1.87 |

PCA revealed contrasts with eigenvalues > 2 for PC and OM (see Table 5.8). However, the contrasts on each test had high disattenuated correlations (.95 and .88, respectively). Therefore, these contrasts are not concerning.

**Table 5.8**

Principal Components Analysis of Standardized Rasch Residuals

| Test | Eigenvalue of Largest Contrast | Variance Explained by Largest Contrast | Variance Explained by Rasch Dimension |
|---|---|---|---|
| Postural Control | 2.24 | 6.8% | 42.7% |
| Balance | 1.86 | 4.4% | 71.4% |
| Ocular Motor | 2.01 | 8.8% | 38.7% |
| Bilateral Integration | 1.62 | 5.9% | 45.7% |

DIF analyses based on sex revealed no items with DIF > .43 logits. Although several items reached statistical significance (marked by asterisks on Tables 5.2-5.5), contrasts were small and therefore likely not meaningful.

Figures 5.1-5.4 show the Winsteps-generated Wright maps for each of the sensory perception tests. Table 5.9 describes the mean person abilities for each test. For PC, items are heavily weighted towards the bottom of the scale, with the child mean (2.74 logits) far exceeding

the item mean (fixed at 0.0 logits). While the items are likely too easy for the normative group, the items are ordered logically. Head lag (a phenomenon which typically disappears in infancy; Flanagan et al., 2012) is the easiest item, followed closely by simple reaching tasks that require relatively little postural control. Supine flexion and prone extension, which require postural control of all body segments against gravity, require more postural control and, thus, fall in the middle of the scale (Assaiante et al., 2005). Children appear to be most challenged by the timed items (PE Time and SF Time), which require them to maintain novel positions for an extended period.

The BAL Wright map suggests that the items are somewhat easy for the sample population; the mean person measure (1.36 logits) was substantially higher than the mean item difficulty. However, they are spread along the construct, ranging from -4.56 to 4.81 logits. I observed redundancy among items 10, 5 and 6. The hierarchy is logical, with items becoming progressively more difficult when the base of support was smaller (i.e., standing on two feet was easier than standing on toes). Further, my findings are logical and consistent with previous literature suggesting that balance with eyes open is easier than balance with eyes closed (e.g., Hammami et al., 2014; Spasic et al., 2022).

The OM Wright map and mean scores demonstrate that items were very easy for the sample population (37.3% of children achieved a perfect score on this assessment, and the mean person measure was 2.90). Quick localization items were generally the easiest, followed by pursuits. Stabilization items, which required children to maintain their eyes in midline while moving their head, were slightly more difficult. This is consistent with existing literature examining ocular motor development (Schubert & Zee, 2010).

The BI Wright map shows adequate spread of items along the map, although mean child measure (1.55 logits) was still markedly higher than the item difficulty mean (0.0 logits). The items are ordered logically, with simpler items such as marching (Item 10) and alternating tapping hands on the thighs (Item 1) at the bottom of the map. More difficult items that require fine motor precision (bilateral sequential finger touching, Item 9) and complex tapping patterns (Item 3) show higher difficulty measures.

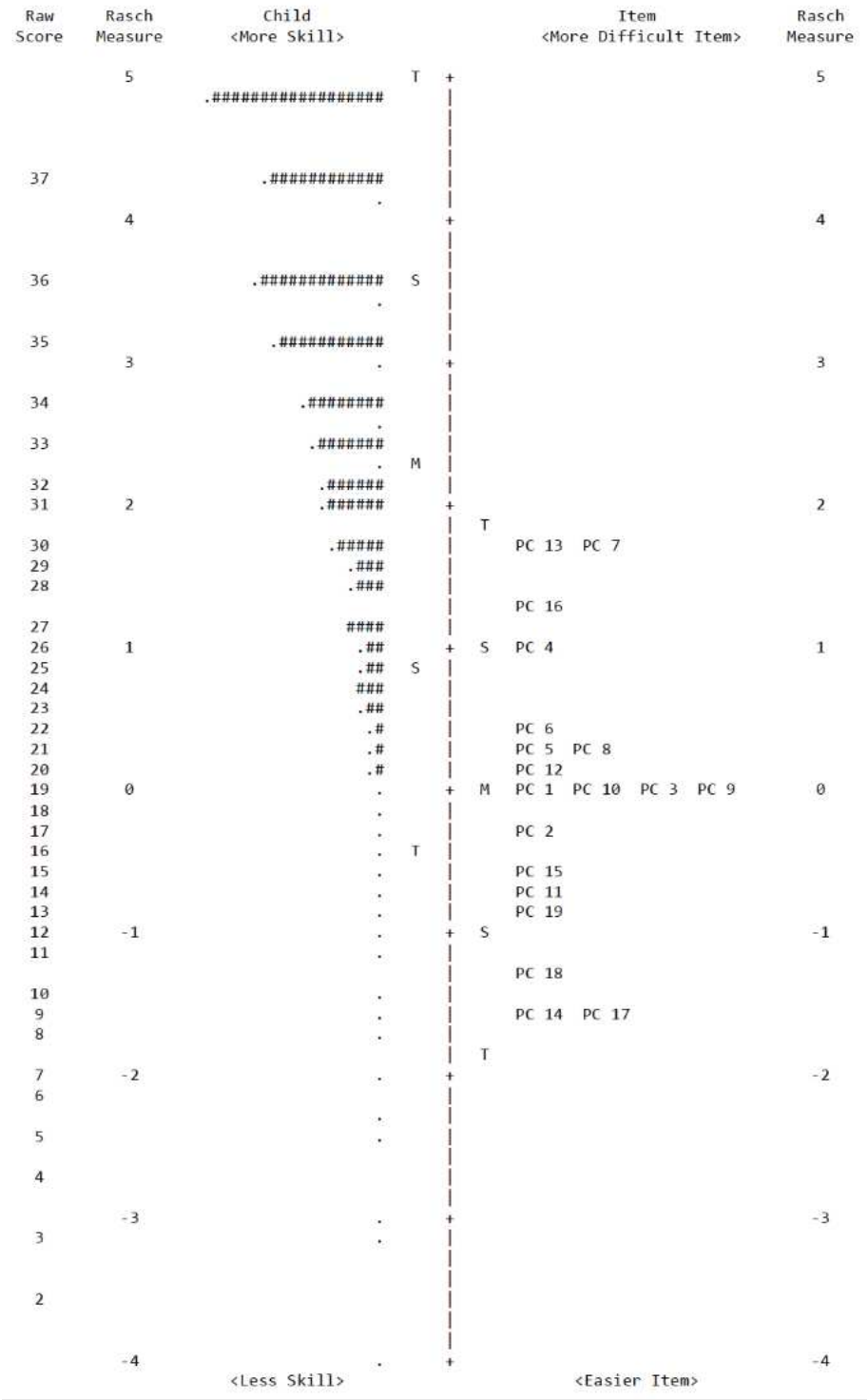



**Figure 5.1**

*Postural Control Wright Map*

*Note*. # = 20 children, . = 1 to 19 children, | = latent trait (postural control), M = item/child mean, S = 1 SD, T = 2 SD

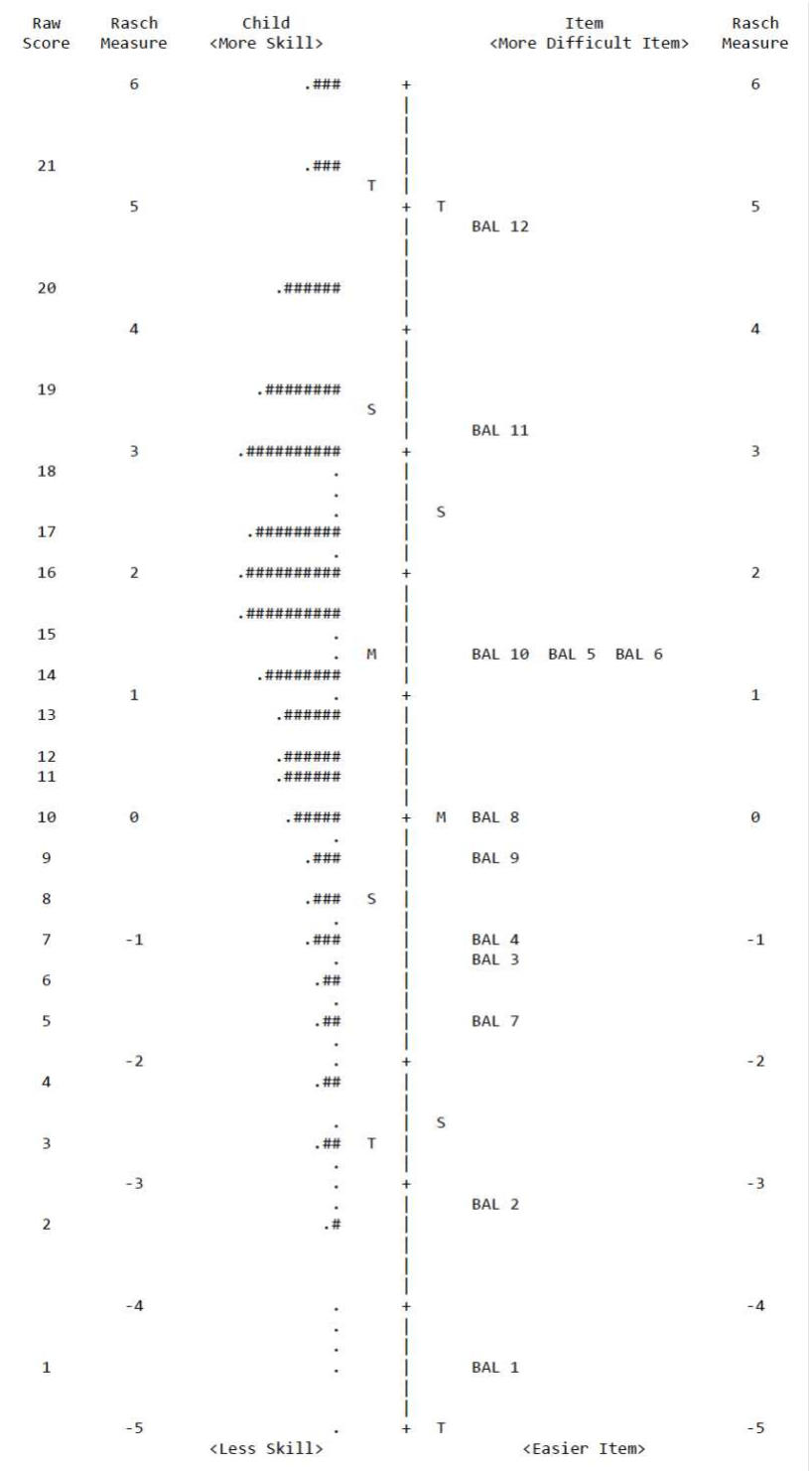



**Figure 5.2**

*Balance Wright Map*

*Note*. # = 20 children, . = 1 to 19 children, | = latent trait (balance), M = item/child mean, S = 1 SD, T = 2 SD

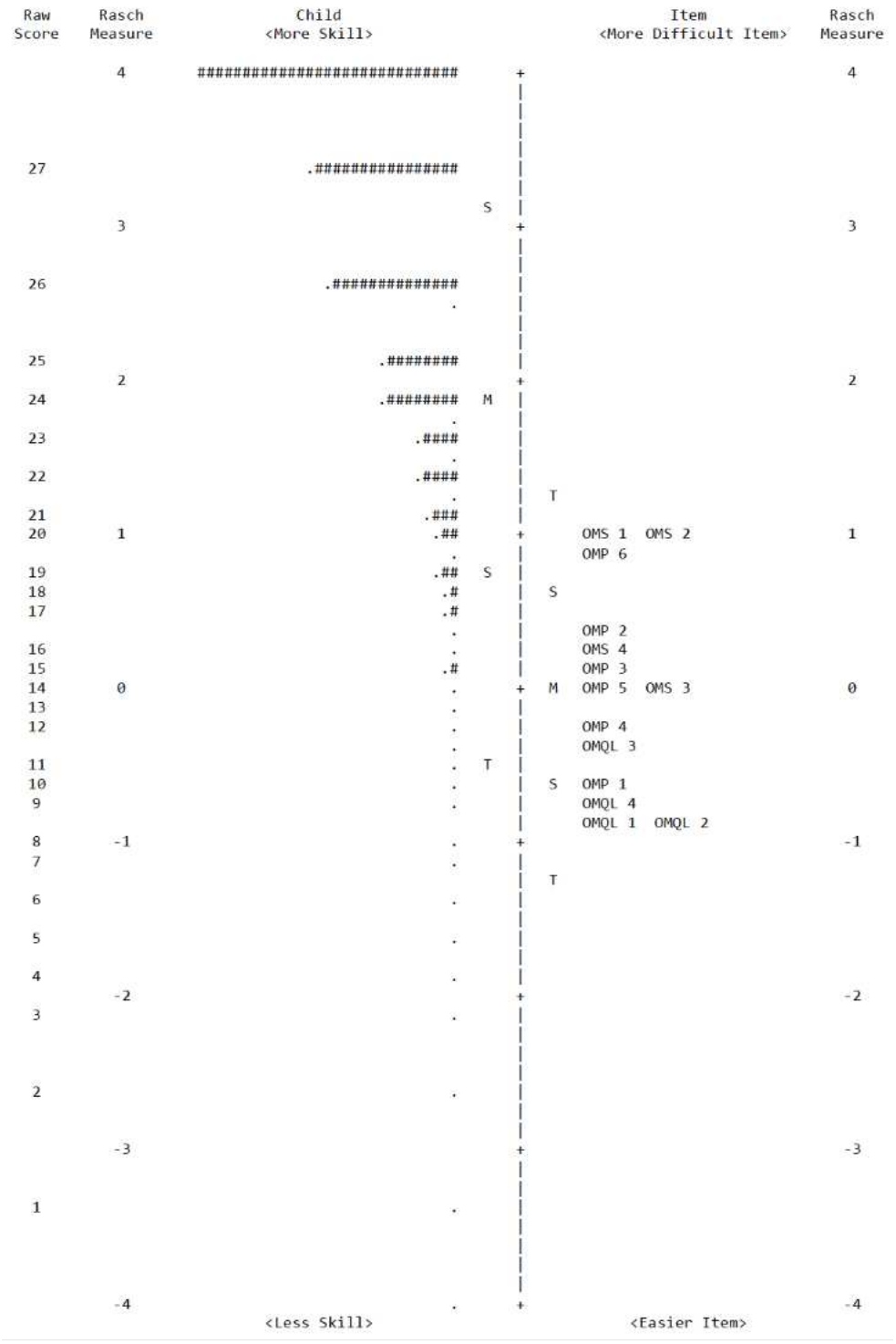



**Figure 5.3**

*Ocular Motor Wright Map*

*Note*. # = 20 children, . = 1 to 19 children, | = latent trait (ocular motor ability), M = item/child mean, S = 1 SD, T = 2 SD

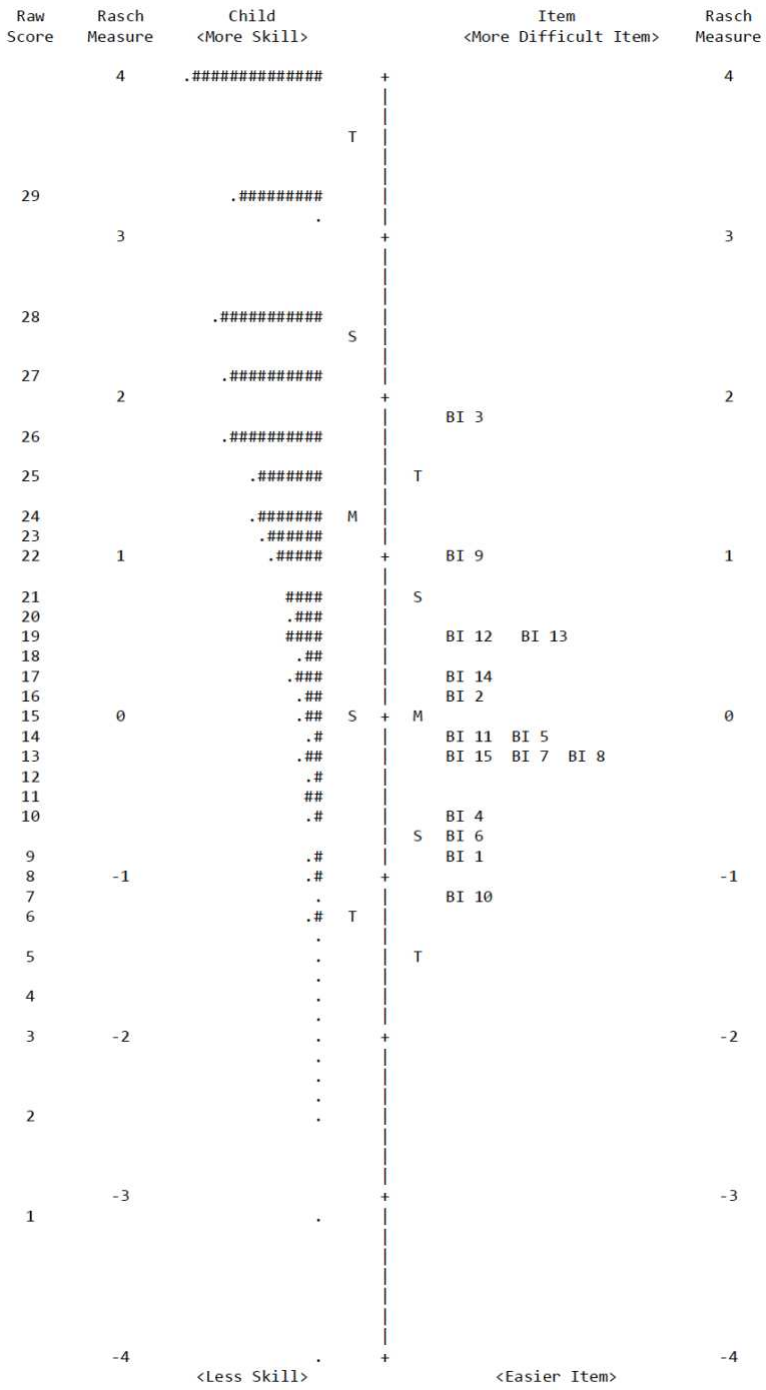



**Figure 5.4**

*Bilateral Integration Localization Wright Map*

*Note*. # = 20 children, . = 1 to 19 children, | = latent trait (bilateral integration), M = item/child mean, S = 1 SD, T = 2 SD

**Table 5.9**

*Mean Child Measure Scores*

| Test | Child Mean (SE) |
|---|---|
| Postural Control | 2.74 (0.77) |
| Balance | 1.36 (0.72) |
| Ocular Motor | 2.90 (1.10) |
| Bilateral Integration | 1.55 (0.69) |

Person maps (Figures 4.5-4.8) and Pearson correlations (Table 5.10) confirmed that scores progress with increasing age. Notably, measure scores for OM failed to meet normality assumptions based on visual inspection of the histogram; the graph revealed significant negative skew as a result of the high overall child measures. I conducted both Pearson and Spearman rank correlations. These did not differ in magnitude or significance; therefore, I reported only the Pearson correlations in Table 5.10.

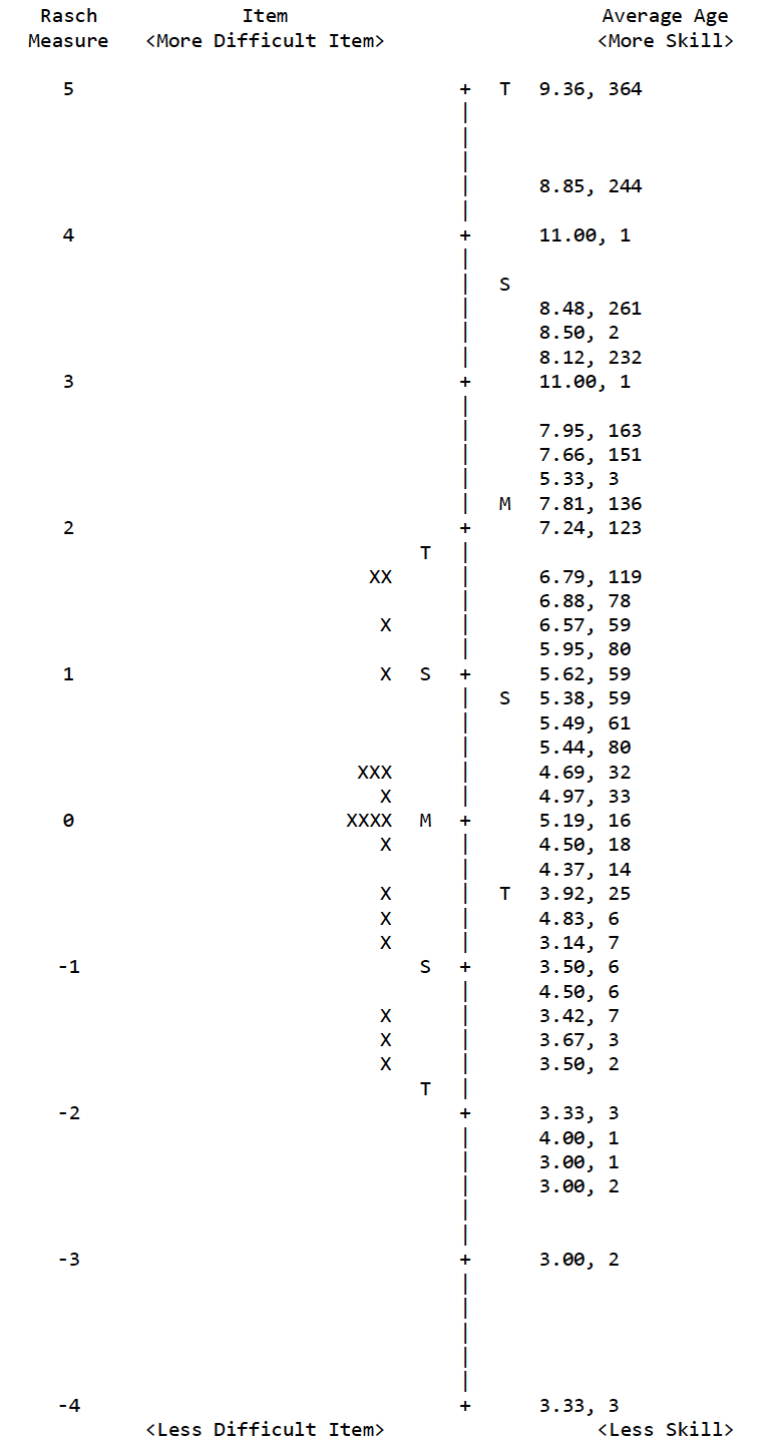



**Figure 5.5**

*Postural Control Person Map*

*Note.* | = latent trait (postural control), left side of the figure shows items (X), right side of the figure shows (average age, # of children), M = item/child mean, S = 1 SD, T = 2 SD

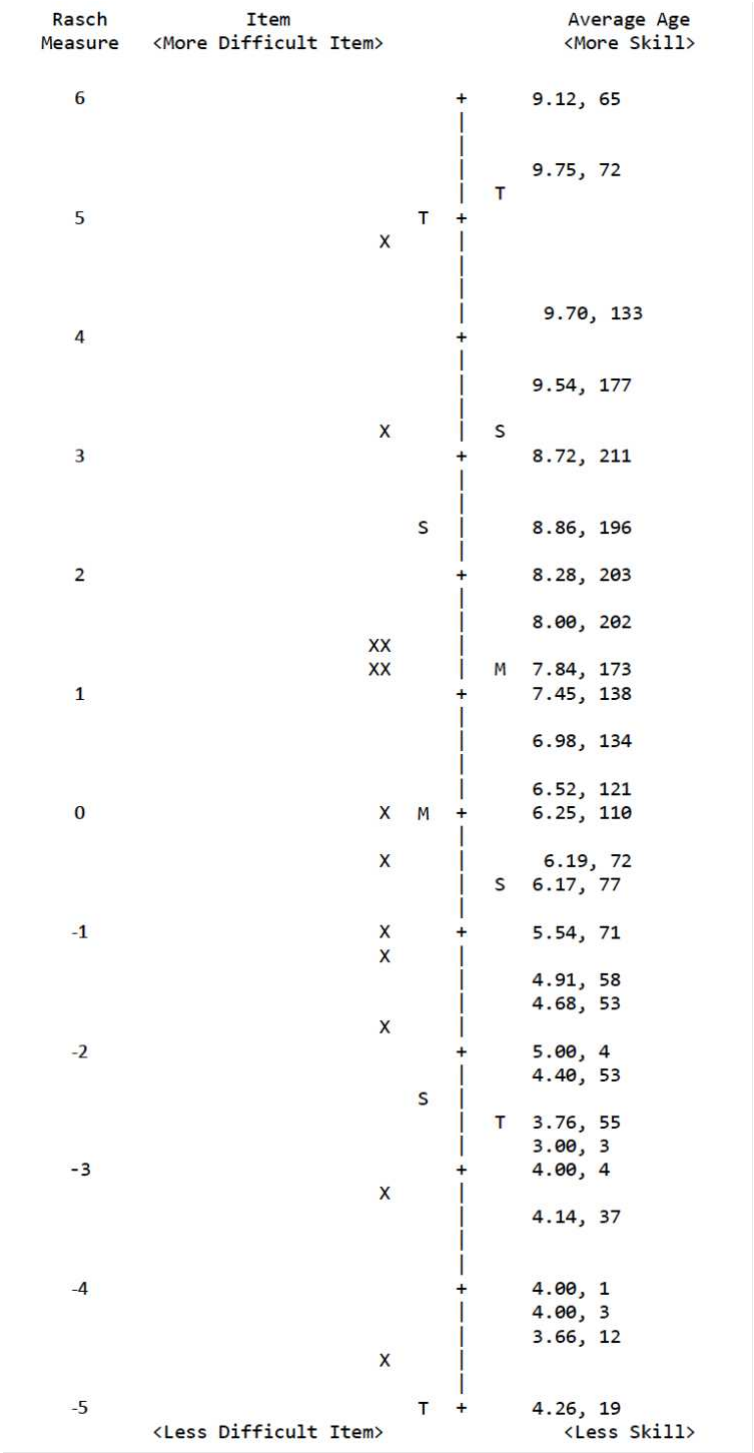



**Figure 5.6**

*Balance Person Map*

*Note.* | = latent trait (balance skill), left side of the figure shows items (X), right side of the figure shows (average age, # of children), M = item/child mean, S = 1 SD, T = 2 SD

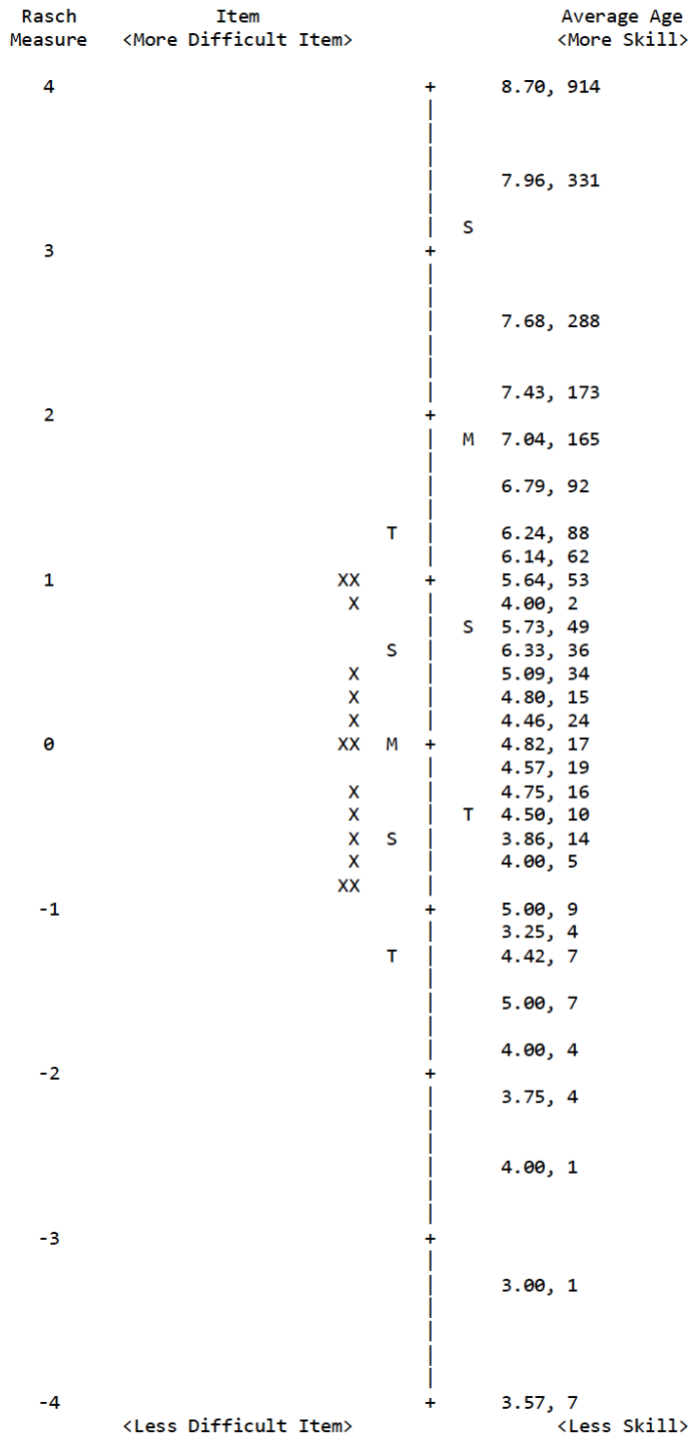



**Figure 5.7**

*Ocular Motor Person Map*

*Note.* | = latent trait (ocular motor ability), left side of the figure shows items (X), right side of the figure shows (average age, # of children), M = item/child mean, S = 1 SD, T = 2 SD

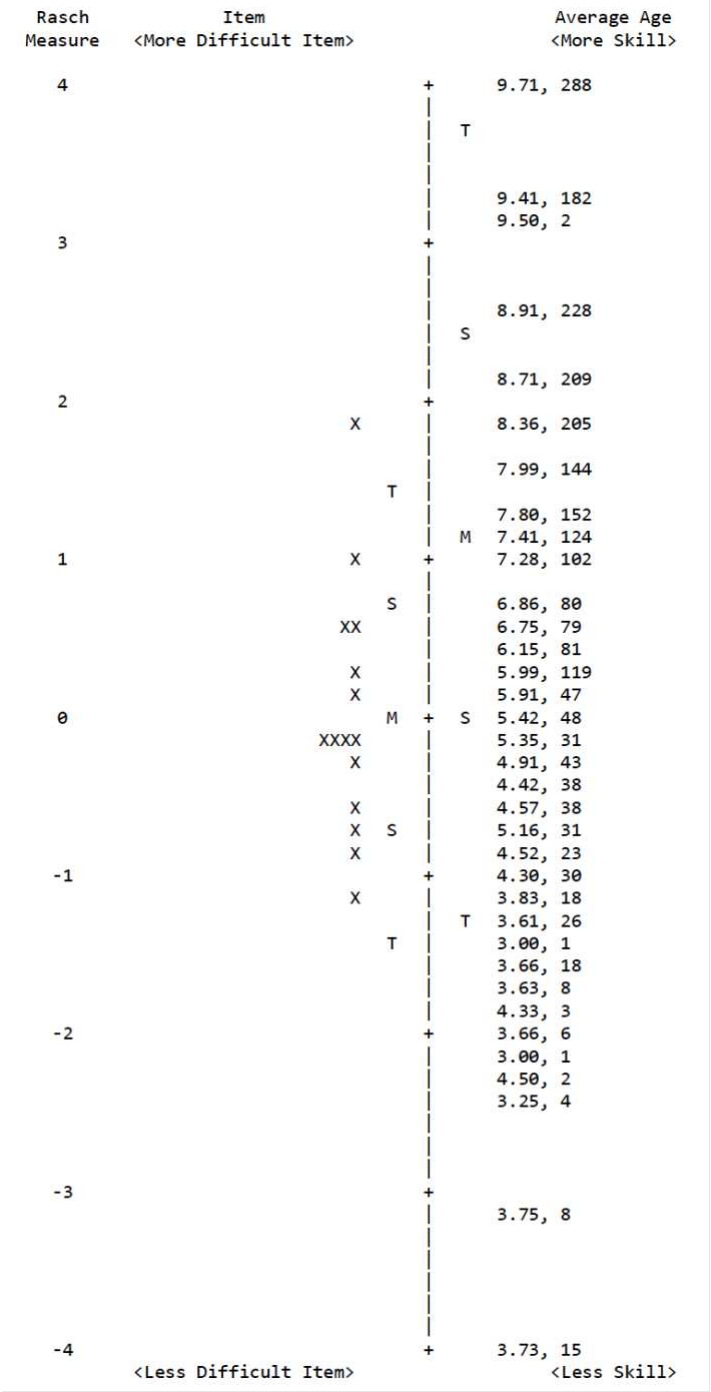



**Figure 5.8**

*Bilateral Integration Person Map*

*Note.* | = latent trait (bilateral integration), left side of the figure shows items (X), right side of the figure shows ((average age, # of children)), M = item/child mean, S = 1 SD, T = 2 SD

**Table 5.10**

*Bivariate Correlations between Age in Months and EASI Motor Measure Scores*

| Test | Correlation |
|---|---|
| Postural Control | 0.56*** |
| Balance | 0.62*** |
| Ocular Motor | 0.49*** |
| Bilateral Integration | 0.65*** |

*Note.* *** = $p < .001$

### Internal Reliability

PC, BAL and BI met our criteria for good/acceptable reliability based on both person reliability indices and strata (See Table 5.11). OM fell slightly below for person reliability index, although strata was acceptable.

**Table 5.11**

*Rasch Reliability*

| Test | Person Reliability Index | Strata |
|---|---|---|
| Postural Control | 0.79 | 2.96 |
| Balance | 0.88 | 3.87 |
| Ocular Motor | 0.66 | 2.20 |
| Bilateral Integration | 0.80 | 2.96 |

## Discussion

The purpose of this study was to evaluate evidence for validity and reliability of the EASI motor tests' normative data using a Rasch analysis approach. Overall, our results are encouraging. Item fit statistics, PCA, and DIF analyses supported the hypothesis that each of these tests measures a unidimensional construct, a critical aspect of construct validity. Rating scale analyses were encouraging. Child fit statistics, visual inspection of the Wright maps, and

person reliability indices suggested that the tests are easy for the sample population; I address this in further detail below.

**Strengths of and Recommended Revisions to the EASI Motor Tests**

The PC, BI, and OM tests all appear relatively easy for the sample population. All of the items on the PC test are motor skills that children should be well on the way to mastering by 3 and 4 years of age; therefore, I expected very high scores among the normative sample. Person maps and correlations support this assertion: scores generally increased with age, but high scores could be achieved by children as young as 6 or 7 years. Similarly, the BI items were generally easy for the normative sample (although less so than PC). Previous results confirm that children with known sensory integration challenges scored significantly lower on both these tests (Mailloux et al., 2021); therefore, these items provide valuable clinical information in the context of sensory integration dysfunction. I suspect that the same theory applies to the OM test; however, no OM results have been published for children with known sensory integration concerns at this time. Currently, I do not recommend changes to OM, BI or PC.

I observed very high outfit MnSq values for BAL 1 and BAL 2. Based on post-hoc inspection of the data, this appears to be because these items are so easy that a handful of low scores conflict with model expectations, resulting in a high MnSq (Linacre, 2002). Our results suggest that even the youngest children can easily complete these tasks; when they could not, this was likely related to inattention, boredom, or failure to understand directions. Therefore, I recommend that these items be converted to trial items – rather than scoring these and artificially inflating total scores, these items should be given to ensure the child's understanding of the task, but not included in the total score.

Notably, two of the four tests (OM and PC) had reliability indices below our threshold for good reliability (.80). PC approached this threshold and showed acceptable internal reliability (.79). OM was lower (.66). These tests also showed the highest discrepancies between child mean and item mean. Linacre (2022) suggested that mismatch between sample ability and item difficulty may be responsible for low reliability. The test developers may improve reliability by adding more difficult items to these tests; however, these items would not add much clinically, as both of these tests are designed to measure skills that most typically developing children will have mastered at young ages. Instead, I expect to find that test reliability is higher for children with sensory integration concerns, for whom these items will be better matched to their level of ability. Researchers should revisit this in future studies.

I suspect that the relative ease of items resulted in lower-than-expected person fit as well. None of the tests reached our threshold of 90%, ranging from 82.7%-89.5%. Person data that fail to fit the Rasch model reflects an unexpected response string (e.g., a child got all difficult items correct, but failed on one or more easy items) (Meijer & Sitsma, 2001). Based on inspection of the data, most children whose data failed to fit the model followed this response pattern across all tests. This is likely due to inattention and/or boredom on items that are far below the ability level of these children.

Although the BI test showed generally adequate evidence for construct validity and reliability, I suspect that BI may be particularly susceptible to influences from praxis. On this test, children are asked to imitate activities (a praxis task in itself) that require smooth and coordinated motions (e.g., tapping their thighs rhythmically). A previous study using the SIPT equivalent to BI, a test called Bilateral Motor Control (BMC), showed that bilateral integration may be better considered a construct of praxis (Lai et al., 1996). Bundy and Lane (2020)

categorized VBIS deficits as a mild form of dyspraxia. VBIS involves difficulty with timing and smooth coordinated motions, especially those using both sides of the body in conjunction. At this time, our results do not suggest any necessary changes to the BI test; however, future researchers should analyze the BI test in conjunction with the EASI praxis tests to determine if this test is a better reflection of motor ability or praxis.

In general, the results of this Rasch analysis suggested that the EASI motor tests provide valid and reliable data for typically developing children. Of particular importance, the item hierarchies are logical, and the person measures increase with increasing age. Additionally, despite the relative ease of the tests, item fit statistics were generally acceptable, suggesting good adherence to the Rasch model. Future researchers should examine the performance of these tests with children who have known sensorimotor concerns; it is likely that construct validity evidence will be even stronger in this population.

**Start and Stop Criteria: A Future Direction for EASI**

Start and stop criteria may increase the clinical feasibility of EASI. Right now, the tests are quite long, which may lead to problems such as boredom and inattention (it is likely that these factors contribute to lower-than-expected person fit statistics). Using Rasch item calibrations, we can establish criteria for flexible start points based on age and order items based on difficulty so that they progress to more difficult items. Children can stop testing when they have missed a certain number of items.

Start and stop criteria are, unfortunately, not a perfect solution. PC, for example, requires testers to evaluate multiple aspects of single tasks (i.e., prone extension and supine flexion). Some may be easy for older children, while others are still valuable to measure. For this test, testers may need to complete all items despite the child's age or ability. Future analyses will

reveal the best solution to create parsimonious EASI tests that maximize both clinical utility and information.

**Developmental Trends Observed in EASI Motor Scores**

Each of the tests demonstrated a strong positive correlation with age. This confirms expectations that the test items form a developmental construct where skills improve as the child ages. Consistent with Ayres's (1972) hypotheses, the items that comprise the PC test were mastered at the youngest ages. The BAL person map shows that children reached the highest levels of performance by approximately age nine, with many children achieving maximum scores by age six or seven. Previous studies and reviews support the notion that balance develops quickly through 7 years of age but continues to develop through adolescence (Assaiante, 1998). Similarly, previous research supports our findings for OM, with rapid improvements in saccadic eye movements through adolescence (Klein et al., 2011).

**Limitations**

Although this study used a large, international sample, examiners recruited children based on convenience; many children were known to the examiners in advance of testing. Also, most examiners were clinicians working with children who have SI disorders. To prevent a skewed sample, our team asked examiners not to recruit siblings of children in their practice, as siblings may be more likely to share diagnoses and may not represent the typically developing population. However, I recognize that this may have 'over-corrected' the sample and resulted in a sample with higher ability than the true typically developing population.

Our normative data collection efforts took place during the COVID-19 pandemic. As a result of varied restrictions across countries, our sample is not evenly distributed across the included world regions. Data collection is still underway. As the restrictions change, more

countries will meet their normative data collection goals, resulting in a more representative population. When sample sizes permit, DIF analyses should be completed by countries and/or world regions to examine whether norms should be stratified by location.

## Conclusions

The results of this study provide evidence for validity and reliability of the normative data collected using the EASI Motor tests (PC, BAL, OM, and BI). These tests are a promising approach to evaluation in sensory integration theory; they build upon previous assessments and can be administered alongside sensory perception and praxis tests. However, these analyses do reveal the relative ease of items for typically-developing children – further studies should investigate their usefulness with children who have SI disorders.

# CHAPTER 6: CONCLUSION

In this dissertation, I examined the evidence for validity and reliability of data collected using 19 of the 21 tests that comprise EASI, using the Rasch model. Overall, most tests met expectations for unidimensionality with acceptable item fit, acceptable or near-acceptable person fit, suitable rating scales, insignificant PCA, and negligible DIF findings. Several tests, however, failed to meet these criteria and therefore require significant revision or very close monitoring during future studies: VPrD, PF and PJP (see Manuscripts 1 and 2 for detailed information and recommendations).

Furthermore, I observed lower-than-expected person fit and reliability across many of the EASI tests. The most likely explanation is that these children were typically developing, and, therefore, many items were quite easy for them. Unexpected errors on very easy items (often caused by inattention or boredom) have outsized impacts on both fit and reliability. However, such small errors are unlikely to impact their overall significantly high scores. In other words, the risk that these children will be incorrectly classified as having SI dysfunction is low.

Despite the simple explanation, these parameters must be examined again in future studies. Notably, previous studies using earlier versions of EASI with children with and without known SI dysfunction found slightly higher person reliability on many of these tests (Mailloux et al., 2021, Schaaf et al., in press, Mailloux et al., under review) - this lends credence to the theory that low reliability can be explained by the ease of items for a typical population of 3-12 year olds. Additional studies must be conducted to evaluate the reliability of scores with children who have SI dysfunction and other diagnoses.

## The EASI: A Solution to Problems in SI Assessment

In Chapter 2, I conducted an extensive review of literature examining SIPT and other alternatives to assessment in SI. However, these assessments presented a variety of challenges including lack of alignment with SI theory, poor or potentially biased psychometric properties, lack of global normative data/standardization samples and difficulties with clinical utility. In many ways, the EASI meets the demand for a novel assessment to evaluate SI.

First and foremost, EASI is aligned with SI theory (Mailloux et al., 2018). Each test measures a unique construct of SI theory, drawing upon years of research conducted by Ayres and her successors. For the most part, this study demonstrated that these are unidimensional constructs which can be well-measured by the EASI tests. Alignment with theory is critical to evaluation in SI (Parham et al., 2011). Instead of piecemeal observations conducted using multiple tests from different theoretical backgrounds (as OTs often must do), therapists using an SI approach with children ages 3-12 now have access to an instrument that is directly suited to their theoretical approach.

Second, this dissertation presents promising evidence that data collected using the EASI is valid and reliable for clinical interpretation. Most tests showed strong evidence for unidimensionality, invariance, and reliability – testers can trust that these scores largely represent the child's SI function. However, several tests require additional revision or investigation to establish validity and reliability – I will discuss this more in the next section. Despite some shortcomings, I am generally confident that EASI tests produce valid and reliable data.

Third, the dataset used in this study represents a geographically, linguistically, socioeconomically, racially and ethnically diverse population. This is in line with the population that receives SI therapy – SI is a global theory with qualified therapists on every continent and in

hundreds of countries (Smith Roley, Mailloux & Erwin, 2022). While our research team must still collect additional data (e.g., with clinical populations), our normative data collection represents a monumental effort to diversify the population for whom EASI is useful.

Fourth, the EASI presents some benefits in terms of clinical utility – but it is far from perfect in this regard. While SIPT were quite costly to purchase and administer, EASI is relatively inexpensive and can be scored for free using the online scoring platform. We have also attempted to remove items that are redundant, shortening the tests substantially. This continued during this dissertation; we removed a handful of items that showed poor fit or added little to clinical interpretation. However, EASI is still quite long. Our research team must continue to shorten these tests to address concerns with utility and feasibility in the clinical setting.

I also noted that very few of the tests often used by therapists evaluating SI have evidence for responsivity. As a result, these tests should not be used as outcome measures for research or clinical practice. Similarly, we have not yet established the strength of this measurement property for EASI. As we move forward, the EASI research team should determine whether the EASI is adequately responsive to change to be used as an outcome measure.

## Future Directions for EASI

In September of 2022, our team enrolled the first several hundred examiners in the EASI Scoring Program (ESP) – an online scoring platform that will provide norm-based scores for clinicians using the EASI in their clinical practice. As a result of these analyses, ESP does not currently provide standardized scores for PF, and the site omits the two-hand items from the PJP test. Both tests are new to the EASI (joint perception was previously measured on KIN on SIPT; however, the two-hand items are new). Researchers should continue to revise/monitor these tests until studies suggest that they produce valid and reliable data for use in clinical populations.

**Future Studies with Clinical Populations**

When considering the future of EASI, it is critical that we keep in mind that validity and reliability are not fixed properties of a given test. Rather, they are situational: a test may be valid and sufficiently reliable for answering a certain question with a certain population (e.g., does this child qualify for services? Does this child have a primary praxis problem or a primary reactivity problem?) (AERA, 2014). The central question that this dissertation examined was: are the EASI tests valid and reliable for producing normative data with typically developing children? While my results suggest that the general answer is "yes", they also emphasize the need for additional research with clinical populations.

Benjamin Wright (one of the main psychometricians behind the Rasch model) always emphasized the importance of studying a measure on the population for whom it is intended (Bundy, A., personal communication, February 2022). In the case of the EASI, the vast majority of users will use this test to examine children who appear to have sensory integration problems: they will examine children referred to sensory integration clinics as a result of behavioral symptoms or occupational problems that may be explained by SI dysfunction. Therefore, it is critically important to re-examine the measurement properties established in this dissertation with children who have known or suspected SI dysfunction.

Additionally, these tests should be examined with groups of children who have other clinical diagnoses. SI dysfunction often occurs concurrently with a variety of other conditions, including ADHD, ASD (Bundy & Lane, 2020), Down Syndrome (Bruni et al., 2010), Fragile X (Rais et al., 2018) and many others. EASI must be validated for use with these populations as well (children with ASD and ADHD were included in previous studies and piloting of EASI; others were excluded). Currently, our studies have also excluded children with IQ < 70 due to the

verbal demands of the EASI – however, because sensory integration concerns often overlap with diagnoses that cause intellectual disability, I suggest that we trial EASI with populations of children with lower IQ.

**Future Studies Examining Cultural/Regional Impacts**

Originally, I intended this dissertation to include DIF analyses of languages, countries and/or global regions. Van Jaarsveld and colleagues' (2012) study suggested that South African children required different normative data than North American children on SIPT. This may or may not be true for EASI as well, given that the EASI development team aimed to keep items free of cultural influence. However, because of the COVID-19 pandemic, examiners (i.e., data collectors) from some countries were unable to submit sufficient scores at the time I retrieved data for this study. Therefore, I omitted this section and considered only sex for DIF analyses. When data collection is complete (target date December 2022), we will conduct these studies to determine whether a single set of international norms is appropriate for measuring all children. Based on preliminary analyses (not included in this dissertation), I am optimistic that these norms will hold up to future DIF analyses.

**Measuring Older Children with SI dysfunction**

One common theme across each analysis is a need for more difficult items suitable for measuring the oldest children with SI dysfunction – it is reasonable to expect that these children will be able to complete many of the items that differentiate younger children with and without SI dysfunction. Although our previous analyses showed that EASI tests could differentiate between children with and without known SI dysfunction, these studies contained few children older than 10. It is possible that these children will not be reliably measured by EASI. If this proves to be the case, some tests may benefit from additional more difficult items.

I hesitate to recommend adding more items because of the length of the existing tests. However, several solutions may allow us to explore additional items without adding significant burden to children and clinicians. First, computer adaptive testing is an option for many of the EASI tests. This is discussed in further detail in each manuscript; for some tests (i.e., VPRD), I acknowledge that computer administration would be impractical. Further, not all testers will have access to computers during evaluation.

Rasch item measures may be used to establish computer adaptive testing parameters, but confirmatory studies should first be conducted to validate the item measures presented here. Additionally, we may consider implementing start and stop criteria for these tests based on age. However, before doing so, we must confirm that the validity of these criteria holds up for children with known SI dysfunction and other clinical diagnoses.

## EASI Tests Omitted from the Dissertation

The original dissertation proposal included two EASI tests that I ultimately excluded from the study: Vestibular Nystagmus (VN) and Sensory Reactivity (SR). I omitted each of these tests because they did not meet criteria for using the Rasch model (details below).

### Vestibular Nystagmus

The VN test evaluates the postrotary nystagmus (PRN) reflex. Briefly, PRN refers to nystagmus that occurs when endolymphatic fluid within the semicircular ear canals is subjected to rotational movement (Mulligan, 2011). Practically, the clinician stimulates this reflex by rotating the child on a spinning board or chair with the neck at 30° of flexion for 10 rotations at approximately 1 rotation/2 seconds. Previous studies show that abnormally short or long duration PRN are associated with low scores on tests that measure vestibular functions such as balance and postural control (Mulligan, 2011).

For the VN test of EASI, the examiner administers six sets of rotations – three clockwise and three counterclockwise – at three intervals across the duration of testing. The examiner then records the duration of the nystagmus reflex in seconds. Clockwise and counterclockwise rotations are averaged for two scores on the VN test. Abnormally high or low nystagmus in either direction signifies potential concerns, respectively. Preliminary analyses of the EASI normative data suggested that the two scores – clockwise and counterclockwise were (as expected) highly correlated. Essentially, this is a single item test. The Rasch model is inappropriate for tests without multiple items; therefore, I excluded it from this dissertation.

**Sensory Reactivity**

During the SR test of EASI, the examiner provides auditory, olfactory, tactile, and movement stimuli to the child and observes his or her response. The examiner assigns a score of 0 (no abnormal response) or 1 (hyper-responsivity). Originally, I expected that this test would be appropriate for Rasch analysis – I could test the hypothesis that, although the stimuli varied across sensory modalities, the entire set of items reflected a construct of sensory reactivity. An unexpectedly high score would suggest hyper-responsivity. However, after gross inspection of the data, I determined that the data would not likely be suitable for Rasch analysis. The vast majority of children showed no hyper-responsivity to any items and would therefore occupy extreme scores in the Rasch model. The model excludes extreme scores, as a child with a perfect score's true ability cannot be adequately reflected by the given items (Wright, 1998). The lack of variation in scores would make item calibration of this test untrustworthy.

In a way, it is encouraging that most children had extreme scores. Hyper-responsivity is, by definition, abnormal. Our team aimed to only include typically developing children in the normative sample. Children who showed hyper-reactivity to any items would be unlikely to be

typically developing, or their responses would suggest that our items were not reflective of the central construct. In the future, we will examine these items in more depth with children who have known or suspected SI dysfunction. We have an existing small sample of children with known SI dysfunction whose scores form a promising Rasch construct (Grady-Dominguez, unpublished data). The results of this analysis were promising; I hope for similar results in larger analyses.

In addition to the discrete SR test, the other EASI tests contain many items for which the examiner can score the child's reactivity. For example, on the TPD tests, the examiner scores not just the child's accuracy in recreating the design, but also their reaction to the sensation on the forearm/hand. Unlike the SR test, where children are scored only on hyper-reactivity, the examiner also notes the presence of hypo-reactivity. It is possible that these items may also contribute to the SR construct, and therefore should be scored together. However, I faced the same problem with these items: abnormal responses were exceedingly rare. Furthermore, we are not confident that testers could accurately distinguish between hypo-reactivity and poor sensory discrimination. In the meantime, we are asking EASI examiners to continue recording these scores; we will re-evaluate their usefulness in future studies.

## Future of this Dissertation

Currently, I have not submitted these three manuscripts for publication. Normative data collection has not been completed as a result of delays related to the COVID-19 pandemic. The EASI research team determined that publishing preliminary results using these data is impractical; instead, we will conduct confirmatory analyses when data collection is complete and update these manuscripts accordingly. Given the large dataset used in this study, the results here

are unlikely to change significantly with the addition of more typically developing children; however, the most practical approach is to publish all the data at once.

## My Future as a Researcher

At the beginning of the Occupational and Rehabilitation Science Ph.D. program, I was enthusiastic about play research – for this reason, I sought out Dr. Bundy as an advisor. However, as I have developed as a researcher, I have found that my passion is truly with measurement and the development of clinically useful assessments that produce valid and reliable data for occupational therapists, researchers and related professionals. While I feel very lucky to have spent the last four years immersed in sensory integration literature, I am ultimately not tied to this population or field of study. The skills that I have developed during my PhD will allow me to work on assessments for a variety of client populations and problems impacting occupational participation and performance.

Regardless of my next direction, my goal is to ensure that useful assessments make their way into routine practice. In their seminal article, "Implementation science: What is it and why should I care?", Bauer and Kirchner (2020) warned that "establishing the effectiveness of a clinical innovation is not sufficient to guarantee its uptake into routine use" (p. 1). Researchers in the field of measure development must take this message to heart: wonderful assessments that produce valid and reliable data are completely useless if they sit on shelves, untouched by clinicians and not impacting clients. Through the process of developing the EASI and the EASI scoring program, I have witnessed the importance of ensuring that an assessment both produces psychometrically strong data *and* is clinically feasible and appropriate.

Immersing myself in the field of implementation science is a logical next step for my career. Implementation science focuses on dissemination, adoption, and maintenance of

evidence-based practices in existing systems, including novel tools for assessment (Bauer & Kirchner, 2020). This field aims to close the gap between research and practice through tested, reliable approaches that fit new approaches into existing workflows. I am seeking postdoctoral opportunities with researchers who are using implementation science to improve uptake of novel assessments.

On a personal level, completion of this dissertation has tested my tenacity, my commitment to research, and my faith in occupational and rehabilitation science as an important field of study. I sought out a career in occupational therapy because I believe in the power of ordinary activities to achieve extraordinary results, form the basis for extraordinary accomplishments, and allow people to live extraordinary lives. For many children, SI dysfunction presents a barrier to ordinary activities. Through targeted, thoughtful, and evidence-based interventions crafted and tailored by skilled occupational therapists, these children can go on to experience occupational successes – ordinary milestones and daily achievements – that allow them to participate within their communities, achieve their dreams, and enjoy more moments in their everyday lives. These interventions rely on valid and reliable data generated by clinically useful assessments. This dissertation forms the basis for the EASI: an assessment that can lead to excellent intervention. Despite the trials I faced in completing this dissertation (most notably, a global respiratory pandemic, several health crises, and balancing multiple caregiver, work, and education responsibilities), these children's ordinary lives motivated me every single day. I submit this dissertation with confidence that it represents a small step towards a better future for children impacted by SI dysfunction.

# REFERENCES


Ahn, R. R., Miller, L. J., Milberger, S., & McIntosh, D. N. (2004). Prevalence of parents' perceptions of sensory processing disorders among kindergarten children. *American Journal of Occupational Therapy*, *58*(3), 287-293.

Alkhalifah, S. M., AlArifi, H., AlHeizan, M., Aldhalaan, H., & Fombonne, E. (2020). Validation of the Arabic Version of the Two Sensory Processing Measure Questionnaires. *Research in Autism Spectrum Disorders*, *78*, 101652.

American Educational Research Association., American Psychological Association., National Council on Measurement in Education., & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Assaiante, C. (1998). Development of locomotor balance control in healthy children. *Neuroscience Biobehavioral Reviews*, *22*(4), 527-532. https://doi.org/10.1016/s0149-7634(97)00040-7

Assaiante, C., Mallau, S., Viel, S., Jover, M., & Schmitz, C. (2005). Development of postural control in healthy children: a functional approach. *Neural Plasticity*, *12*(2-3), 109-118; discussion 263-172. https://doi.org/10.1155/NP.2005.109

Ausderau, K., Sideris, J., Furlong, M., Little, L. M., Bulluck, J., & Baranek, G. T. (2014). National survey of sensory features in children with ASD: Factor structure of the sensory experience questionnaire (3.0). *Journal of autism and developmental disorders*, *44*(4), 915-925.

Ayres, A. J. (1972). *Sensory integration and learning disorders*. Western Psychological Services (WPS).

Ayres, A. J. (1989). Sensory Integration and Praxis Tests manual. Los Angeles: Western Psychological Services.

Ayres, A. J. (1991). Sensory Integration and Praxis Tests (SIPT). In Torrance, CA: Western Psychological Services (WPS).

Ayres, A. J. (2005). *Sensory integration and the child: Understanding hidden sensory challenges*. Western Psychological Services (WPS).

Ayres, A. J. (2014). Sensory integration and praxis tests (SIPT). In (4th ed.). Torrance, CA: Western Psychological Services (WPS).

Ayres, A. J., & Robbins, J. (1979). Sensory integration and the child. Torrance, CA: Western Psychological Services (WPS).

Barnett, L. A. (1991). The playful child: measurement of a disposition to play. *Play & Culture*, *4*(1), 51-74.

Bauer, M. S., & Kirchner, J. (2020). Implementation science: What is it and why should I care? Psychiatry Res, 283, 112376. https://doi.org/10.1016/j.psychres.2019.04.025

Beery, K. (2010). Beery VMI administration, scoring, and teaching manual. *Bloomington, MN: PsychCorp*.

Beery, K., Buktenica, N., & Beery, N. (1967). Beery Developmental Test of Visual-Motor Integration (VMI). NCS Pearson.

Ben-Sasson, A., Carter, A. S., & Briggs-Gowan, M. J. (2009). Sensory over-responsivity in elementary school: prevalence and social-emotional correlates. *Journal of abnormal child psychology*, *37*(5), 705-716.

Bezrukikh, M. M., & Terebova, N. N. (2009). Characteristics of the development of visual perception in five- to seven-year-old children. Fiziol Cheloveka, 35(6), 37-42. https://www.ncbi.nlm.nih.gov/pubmed/20063705

Blanche, E. I., Bodison, S., Chang, M. C., & Reinoso, G. (2012). Development of the Comprehensive Observations of Proprioception (COP): Validity, reliability, and factor analysis. *American Journal of Occupational Therapy*, *66*(6), 691-698.

Blanche, E. I., Reinoso, G., & Kiefer, D. B. (2021). Structured observations of sensory integration-motor (SOSI-M) & comprehensive observations of proprioception (COP-R). Western Psychological Services.

Blanche, E., Reinoso, G., Kiefer, D.B., (2020). Using Clinical Observation with the Evaluation Process. In *Sensory Integration: Theory and Practice* (pp. 222-241). F. A. Davis.

Bond, T. G., Yan, Z., & Heene, M. (2020). *Applying the Rasch model : fundamental measurement in the human sciences* (Fourth edition. ed.). Routledge.

Bremner, A. J., & Spence, C. (2017). The Development of Tactile Perception. *Adv Child Dev Behav*, *52*, 227-268. https://doi.org/10.1016/bs.acdb.2016.12.002

Brown, T. (2010). Construct validity: A unitary concept for occupational therapy assessment and measurement. *Hong Kong Journal of Occupational Therapy*, *20*(1), 30-42.

Brown, T. (2016). Validity and reliability of the developmental test of visual perception–third edition (DTVP-3). *Occupational therapy in health care*, *30*(3), 272-287.

Brown, T. (2019). Structural validity of the Bruininks-Oseretsky test of motor proficiency–second edition brief form (BOT-2-BF). *Research in Developmental Disabilities*, *85*, 92-103.

Brown, T., Morrison, I. C., & Stagnitti, K. (2010). The reliability of two sensory processing scales used with school-age children: Comparing the response consistency of mothers, fathers, and classroom teachers rating the same child. *Journal of Occupational Therapy, Schools, & Early Intervention*, *3*(4), 331-347.

Brown, T., Morrison, I. C., & Stagnitti, K. (2010). The convergent validity of two sensory processing scales used with school-age children: comparing the Sensory Profile and the Sensory Processing Measure. *New Zealand journal of occupational therapy*, *57*(2), 56-65.

Brown, T., Unsworth, C., & Lyons, C. (2009). An evaluation of the construct validity of the Developmental Test of Visual-Motor Integration using the Rasch Measurement Model. Australian Occupational Therapy Journal, 56(6), 393-402. https://doi.org/10.1111/j.1440-1630.2009.00811.x

Bruininks, R. H., & Bruininks, B. D. (2005). *Bruininks-Oseretsky Test of Motor Proficiency - Second Edition*. Pearson Assessments.

Bruni, M., Cameron, D., Dua, S., & Noy, S. (2010). Reported sensory processing of children with Down syndrome. *Physical & Occupational Therapy in Pediatrics*, *30*(4), 280-293. https://doi.org/10.3109/01942638.2010.486962

Bundy, A. C., & Lane, S. J. (2020a). *Sensory Integration: Theory and Practice* (3rd ed.). F. A. Davis.

Bundy, A. C., & Lane, S. J. (2020b). Sensory Integration: A. Jean Ayres' Theory Revisited. In *Sensory Integration: Theory and Practice* (3rd ed., pp. 2 - 20). F. A. Davis.

Bundy, A. C., & Szklut, S. (2020). The science of intervention: creating direct intervention from theory. In A. C. Bundy & S. J. Lane (Eds.), Sensory Integration: Theory and Practice (pp. 300-337). F. A. Davis.

Capistrano, R., Ferrari, E. P., Souza, L. P. d., Beltrame, T. S., & Cardoso, F. L. (2015). Concurrent validation of the MABC-2 motor tests and MABC-2 checklist according to the developmental coordination disorder questionnaire-br. *Motriz: Revista de Educação Física*, *21*(1), 100-106.

Case-Smith, J., & O'Brien, J. C. (2014). Occupational therapy for children and adolescents (Seventh edition. ed.). Elsevier.

Cella, D., Gershon, R., Lai, J. S., & Choi, S. (2007). The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment. *Qual Life Res*, *16 Suppl 1*, 133-141. https://doi.org/10.1007/s11136-007-9204-6

Cermak, S. A., & May-Benson, T. A. (2020). Praxis and dyspraxia. In A. Bundy & S. J. Lane (Eds.), Sensory Integration Theory and Practice (pp. 115-150). F.A. Davis.

Cermak, S. A., & Murray, E. A. (1991). The validity of the Constructional subtests of the Sensory Integration and Praxis Tests. American Journal of Occupational Therapy, 45(6), 539-543. https://doi.org/10.5014/ajot.45.6.539

Chojnicka, I., & Pisula, E. (2019). Adaptation and psychometric properties of the Polish version of the Short Sensory Profile 2. *Medicine*, *98*(44).

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). L. Erlbaum Associates. Publisher description http://www.loc.gov/catdir/enhancements/fy0731/88012110-d.html

Colarusso, R. P., & Hammill, D. D. (1972). *Motor-free visual perception test*. ATP Assessments.

D., W. B. (1998). *Model selection: Rating Scale Model (RSM) or Partial Credit Model (PCM)?* Retrieved Oct 4 from www.rasch.org/rmt/rmt1231.htm

Colarusso, R. P., & Hammill, D. D. (2015). Motor-Free Visual Perception Test - Fourth Edition

D., W. B. (1998). Model selection: Rating Scale Model (RSM) or Partial Credit Model (PCM)? Retrieved Oct 4 from www.rasch.org/rmt/rmt1231.htm

Darr, N., Franjoine, M. R., Campbell, S. K., & Smith, E. (2015). Psychometric properties of the pediatric balance scale using rasch analysis. *Pediatric physical therapy*, *27*(4), 337-348.

DeGangi, G. A. (1983). *DeGangi-Berk test of sensory integration (TSI)*. Western Psychological Services.

Diemand, S. C., & Case-Smith J. (2013). *Validity of the Miller Function and Participation Scales* The Ohio State University].

Dua, K., Lancaster, T. P., & Abzug, J. M. (2019). Age-dependent Reliability of Semmes-Weinstein and 2-Point Discrimination Tests in Children. Journal of Pediatric Orthopedics, 39(2), 98-103. https://doi.org/10.1097/BPO.0000000000000892

Dunn, W. (2014). *Sensory Profile 2 Users Manual*. Pearson.

Ellinoudis, T., Evaggelinou, C., Kourtessis, T., Konstantinidou, Z., Venetsanou, F., & Kambas, A. (2011). Reliability and validity of age band 1 of the Movement Assessment Battery for Children–Second Edition. *Research in Developmental Disabilities*, *32*(3), 1046-1051. https://doi.org/10.1016/j.ridd.2011.01.035

Estevan, I., Molina-García, J., Queralt, A., Álvarez, O., Castillo, I., & Barnett, L. (2017). Validity and reliability of the Spanish version of the test of gross motor development–3. *Journal of Motor Learning and Development*, *5*(1), 69-81.

Flanagan, J. E., Landa, R., Bhat, A., & Bauman, M. (2012). Head lag in infants at risk for autism: a preliminary study. *Am J Occup Ther*, *66*(5), 577-585. https://doi.org/10.5014/ajot.2012.004192

Fleurkens-Peeters, M., Janssen, A., Akkermans, R., Zijlmans, W., & Nijhuis-van der Sanden, M. (2019). Movement Assessment Battery for Children-2 (MABC-2): A Cross-Cultural Comparison for Surinamese Children at 5 Years of Age.

Folio, R., & Fewell, R. R. (2000). PDMS-2 Examiner's Manual. In: Austin, TX: Pro-Ed.

Fox, C. M., & Jones, J. A. (1998). Uses of Rasch modeling in counseling psychology research. *Journal of Counseling Psychology*, *45*(1), 30.

Franjoine, M. R., Gunther, J. S., & Taylor, M. J. (2003). Pediatric balance scale: a modified version of the berg balance scale for the school-age child with mild to moderate motor impairment. *Pediatric physical therapy*, *15*(2), 114-128.

Gershon, R. C. (2005). Computer adaptive testing. *J Appl Meas*, *6*(1), 109-127. https://www.ncbi.nlm.nih.gov/pubmed/15701948

Grady-Dominguez, P., Bundy, A., Ragen, J., Wyver, S., Villeneuve, M., Naughton, G., Tranter, P., Eakman, A., Hepburn, S., & Beetham, K. (2019). An observation-based instrument to measure what children with disabilities do on the playground: a Rasch analysis. *International Journal of Play*, *8*(1), 79-93. https://doi.org/10.1080/21594937.2019.1580340

Grady-Dominguez, P., Ihrig, K., J., L. S., Aberle, J., Beetham, K., Ragen, J., Spencer, G., Sterman, J., Tranter, P., Wyver, S., & Bundy, A. (2020). Chapter One - Reframing risk: Working with caregivers of children with disabilities to promote risk-taking in play.

*International Review of Research in Developmental Disabilities*, *59*, 1-45. https://doi.org/10.1016/bs.irrdd.2020.09.001

Greenslade, K. J., & Coggins, T. E. (2016). Brief report: an independent replication and extension of psychometric evidence supporting the theory of mind inventory. *Journal of autism and developmental disorders*, *46*(8), 2785-2790.

Griffiths, A., Morgan, P., Anderson, P. J., Doyle, L. W., Lee, K. J., & Spittle, A. J. (2017). Predictive value of the Movement Assessment Battery for Children-Second Edition at 4 years, for motor impairment at 8 years in children born preterm. *Developmental Medicine & Child Neurology*, *59*(5), 490-496.

Griffiths, A., Toovey, R., Morgan, P. E., & Spittle, A. J. (2018). Psychometric properties of gross motor assessment tools for children: a systematic review. *BMJ open*, *8*(10), e021734.

Haley, S. M., Coster, W. J., Dumas, H. M., Fragala-Pinkham, M. A., Kramer, J., Ni, P., Tian, F., Kao, Y. C., Moed, R., & Ludlow, L. H. (2011). Accuracy and precision of the Pediatric Evaluation of Disability Inventory computer-adaptive tests (PEDI-CAT). *Dev Med Child Neurol*, *53*(12), 1100-1106. https://doi.org/10.1111/j.1469-8749.2011.04107.x

Hall, L., & Case-Smith, J. (2007). The effect of sound-based intervention on children with sensory processing disorders and visual-motor delays. American Journal of Occupational Therapy, 61(2), 209-215. https://doi.org/10.5014/ajot.61.2.209

Hambleton, R. K., & Jones, R. W. (1993). Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development In *ITEMS. Instructional Topics in Educational Measurement* (pp. 38-47). National Council on Measurement in Education.

Hammami, R., Behm, D. G., Chtara, M., Ben Othman, A., & Chaouachi, A. (2014). Comparison of static balance and the role of vision in elite athletes. *J Hum Kinet*, *41*, 33-41. https://doi.org/10.2478/hukin-2014-0030

Hammill, D. D., Pearson, N. A., & Voress, J. K. (2014). Developmental Test of Visual Perception (3rd ed.). PRO-ED.

Hansen, K. D., & Jirikowic, T. (2013). A comparison of the sensory profile and sensory processing measure home form for children with fetal alcohol spectrum disorders. *Physical & occupational therapy in pediatrics*, *33*(4), 440-452.

Harvey, E. M., Leonard-Green, T. K., Mohan, K. M., Kulp, M. T., Davis, A. L., Miller, J. M., Twelker, J. D., Campus, I., & Dennis, L. K. (2017). Inter-Rater and Test-Retest Reliability of the Beery VMI in Schoolchildren. *Optometry and vision science: official publication of the American Academy of Optometry*, *94*(5), 598.

Hatzfeld, C., Kern, T. A., & Werthschützky, R. (2010). Design and evaluation of a measuring system for human force perception parameters. *Sensors and Actuators A: Physical*, *162*(2), 202-209. https://doi.org/10.1016/j.sna.2010.01.026

Henderson, S. E., Sugden, D. A., & Barnett, A. L. (2012). *MABC-2: Movement Assessment Battery for Children - 2nd Edition*. PsychCorp.

Hilz, M. J., Axelrod, F. B., Hermann, K., Haertl, U., Duetsch, M., & Neundorfer, B. (1998). Normative values of vibratory perception in 530 children, juveniles and adults aged 3-79 years. J Neurol Sci, 159(2), 219-225. https://doi.org/10.1016/s0022-510x(98)00177-4

Hirata, S., Kita, Y., Yasunaga, M., Suzuki, K., Okumura, Y., Okuzumi, H., Hosobuchi, T., Kokubun, M., Inagaki, M., & Nakai, A. (2018). Applicability of the Movement Assessment Battery for Children-(MABC-2) for Japanese children aged 3–6 years: a

preliminary investigation emphasizing internal consistency and factorial validity. *Frontiers in psychology*, *9*, 1452.

Hoehn, T. P., & Baumeister, A. A. (1994). A critique of the application of sensory integration therapy to children with learning disabilities. *Journal of learning disabilities*, *27*(6), 338-350.

Holloway, J. M., Long, T., & Biasini, F. (2019). Concurrent Validity of Two Standardized Measures of Gross Motor Function in Young Children with Autism Spectrum Disorder. Phys Occup Ther Pediatr, 39(2), 193-203. https://doi.org/10.1080/01942638.2018.1432006

Holm, I., Tveter, A. T., Aulie, V. S., & Stuge, B. (2013). High intra-and inter-rater chance variation of the movement assessment battery for children 2, ageband 2. *Research in Developmental Disabilities*, *34*(2), 795-800.

Hua, J., Gu, G., Meng, W., & Wu, Z. (2013). Age band 1 of the Movement Assessment Battery for Children-: exploring its usefulness in mainland China. *Research in Developmental Disabilities*, *34*(2), 801-808.

Hwang, J. L., Nochajski, S. M., Linn, R. T., & Wu, Y. W. (2004). The development of the School Function Assessment Chinese version for cross-cultural use in Taiwan. Occup Ther Int, 11(1), 26-39. https://doi.org/10.1002/oti.195

Izadi-Najafabadi, S., Ryan, N., Ghafooripoor, G., Gill, K., & Zwicker, J. G. (2019). Participation of children with developmental coordination disorder. *Res Dev Disabil*, *84*, 75-84. https://doi.org/10.1016/j.ridd.2018.05.011

Johnson, S. P. (2010). Development of visual perception. *Wiley Interdiscip Rev Cogn Sci*, *2*(5), 515-528. https://doi.org/10.1002/wcs.128

Kaplan, F. S., Nixon, J. E., Reitz, M., Rindfleish, L., & Tucker, J. (1985). Age-related changes in proprioception and sensation of joint position. Acta Orthop Scand, 56(1), 72-74. https://doi.org/10.3109/17453678508992984

Kay, L. G., Bundy, A. C., & Clemson, L. M. (2009). Predicting fitness to drive in people with cognitive impairments by using DriveSafe and DriveAware. *Arch Phys Med Rehabil*, *90*(9), 1514-1522. https://doi.org/10.1016/j.apmr.2009.03.011

Keogh, J. & Sugden, D. (1985). *Movement skill development*. New, NY: Macmillan.

Kilroy, E., Aziz-Zadeh, L., & Cermak, S. (2019). Ayres Theories of Autism and Sensory Integration Revisited: What Contemporary Neuroscience Has to Say. Brain Sci, 9(3). https://doi.org/10.3390/brainsci9030068

Kim, H. Y. (2013). Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. *Restor Dent Endod*, *38*(1), 52-54. https://doi.org/10.5395/rde.2013.38.1.52

Kimball, J. G. (1990). Using the Sensory Integration and Praxis Tests to measure change: A pilot study. *American Journal of Occupational Therapy*, *44*(7), 603-608.

King, B. R., Kagerer, F. A., Harring, J. R., Contreras-Vidal, J. L., & Clark, J. E. (2011). Multisensory adaptation of spatial-to-motor transformations in children with developmental coordination disorder. *Exp Brain Res*, *212*(2), 257-265. https://doi.org/10.1007/s00221-011-2722-z

Kita, Y., Suzuki, K., Hirata, S., Sakihara, K., Inagaki, M., & Nakai, A. (2016). Applicability of the Movement Assessment Battery for Children-to Japanese children: A study of the Age Band 2. *Brain and Development*, *38*(8), 706-713.

Klein, C., Rauh, R., & Biscaldi, M. (2011). Patterns of change in ocular motor development. *Exp Brain Res*, *210*(1), 33-44. https://doi.org/10.1007/s00221-011-2601-7

Korkman, M., Kirk, U., & Kemp, S. (2007). *NEPSY II: Clinical and interpretive manual*. Harcourt Assessment, PsychCorp.

Köse, B., Karabulut, E., & Akı, E. (2019). Investigating the interchangeability and clinical utility of MVPT-3 and MVPT-4 for 7–10 year children with and without specific learning disabilities. *Applied Neuropsychology: Child*, 1-8.

Kuhnle, S., Ludwig, A. A., Meuret, S., Kuttner, C., Witte, C., Scholbach, J., Fuchs, M., & Rubsamen, R. (2013). Development of auditory localization accuracy and auditory spatial discrimination in children and adolescents. Audiology and Neurotology, 18(1), 48-62. https://doi.org/10.1159/000342904

Lai, C. Y., Chung, J. C., Chan, C. C., & Li-Tsang, C. W. (2011). Sensory processing measure-HK Chinese version: psychometric properties and pattern of response across environments. *Research in Developmental Disabilities*, *32*(6), 2636-2643.

Lai, J., Fisher, A., Magalhães, L., & Bundy, A. (1996). Construct Validity of the Sensory Integration and Praxis Tests. The Occupational Therapy Journal of Research, 16(2), 75-97. https://doi.org/doi:10.1177/153944929601600201

Lamash, L., Grady-Dominguez, P., Mailloux, Z., Parham, L. D., Schaaf, R. C., Smith Roley, S., & Gal, E. (2022). EASI Praxis Tests: Age Trends and Internal Consistency. *American Journal of Occupational Therapy*, *76*(2). https://doi.org/10.5014/ajot.2022.049145

Lane, H., & Brown, T. (2015). Convergent validity of two motor skill tests used to assess school-age children. *Scandinavian Journal of Occupational Therapy*, *22*(3), 161-172.

Lane, S. J. (2020). Structure and Function of the Sensory Systems. In A. C. Bundy & S. J. Lane (Eds.), *Sensory Integration: Theory and Practice* (3rd ed.). F.A. Davis.

Lane, S. J., Ivey, C. K., & May-Benson, T. A. (2014). Test of Ideational Praxis (TIP): Preliminary findings and interrater and test–retest reliability with preschoolers. *American Journal of Occupational Therapy*, *68*(5), 555-561.

Lane, S. J., Mailloux, Z., Schoen, S., Bundy, A., May-Benson, T. A., Parham, L. D., Smith Roley, S., & Schaaf, R. C. (2019). Neural Foundations of Ayres Sensory Integration((R)). Brain Sci, 9(7). https://doi.org/10.3390/brainsci9070153

Lane, S. J., & Reynolds, S. (2020). Sensory discrimination function and disorders. In A. Bundy & S. J. Lane (Eds.), Sensory Integration Theory and Practice (pp. 181-205).

Leong, H., Carter, M., & Stephenson, J. (2015). Systematic review of sensory integration therapy for individuals with disabilities: Single case design studies. *Research in Developmental Disabilities*, *47*, 334-351.

Linacre, J. M. (2000). Computer-Adaptive Testing: A Methodology Whose Time Has Come. In U. Kang, E. Jean, & J. M. Linacre (Eds.), *Development of Computerised Middle School Achievement Tests* (Vol. 69, pp. 1-58). MESA.

Linacre, J. M. (2000). *Comparing and Choosing between "Partial Credit Models" (PCM) and "Rating Scale Models" (RSM)*. Retrieved Oct 5 from https://www.rasch.org/rmt/rmt143k.htm

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*(2), 878.

Linacre, J. M. (2015). *Fit diagnosis: infit outfit mean-square standardized*. Retrieved Oct 4 from https://www.winsteps.com/winman/misfitdiagnosis.htm

Linacre, J. M. (2018). *Detecting multidimensionality in Rasch data using Winsteps Table 23*. Retrieved April 1 from https://www.youtube.com/watch?v=sna19QemE50&t=3s

Linacre, J. M. (2019). *Rasch Thresholds and Category Frequencies: what happens when Andrich Thresholds become disordered?* Retrieved Sep 30 from https://www.youtube.com/watch?v=Rs3F7a6I8_0&t=6s

Linacre, J. M. (2020). *DIF - DPF - bias - interactions concepts*. Retrieved April 3 from https://www.winsteps.com/winman/difconcepts.htm

Linacre, J. M. (2020). *Differential item functioning DIF pairwise*. Retrieved Sep 30 from

Linacre, J. M. (2021). *Andrich Thresholds: Disordered rating or partial credit structures (Categories or Andrich thresholds)*. Retrieved April 1 from https://www.winsteps.com/winman/disorder.htm

Linacre, J. M. (2022). *Reliability and separation of measures*. Retrieved April 21st from https://www.winsteps.com/winman/reliability.htm

Maeng, H., Webster, E. K., Pitchford, E. A., & Ulrich, D. A. (2017). Inter-and intrarater reliabilities of the test of gross motor development—third edition among experienced TGMD-2 raters. *Adapted Physical Activity Quarterly*, *34*(4), 442-455.

Magalhaes, L. C., Cardoso, A. A., & Missiuna, C. (2011). Activities and participation in children with developmental coordination disorder: a systematic review. *Res Dev Disabil*, *32*(4), 1309-1316. https://doi.org/10.1016/j.ridd.2011.01.029

Magistro, D., Piumatti, G., Carlevaro, F., Sherar, L. B., Esliger, D. W., Bardaglio, G., Magno, F., Zecca, M., & Musella, G. (2018). Measurement invariance of TGMD-3 in children with and without mental and behavioral disorders. *Psychological assessment*, *30*(11), 1421.

Mailloux, Z., Grady-Dominguez, P., Petersen, J., Parham, L. D., Roley, S. S., Bundy, A., & Schaaf, R. C. (2021). Evaluation in Ayres Sensory Integration(R) (EASI) Vestibular and Proprioceptive Tests: Construct Validity and Internal Reliability. *American Journal of Occupational Therapy*, *75*(6). https://doi.org/10.5014/ajot.2021.043166

Mailloux, Z., Mulligan, S., Roley, S. S., Blanche, E., Cermak, S., Coleman, G. G., Bodison, S., & Lane, C. J. (2011). Verification and clarification of patterns of sensory integrative dysfunction. Am J Occup Ther, 65(2), 143-151. https://doi.org/10.5014/ajot.2011.000752

Mailloux, Z., Parham, L. D., Roley, S. S., Ruzzano, L., & Schaaf, R. C. (2018). Introduction to the evaluation in ayres sensory integration®(EASI). *American Journal of Occupational Therapy*, *72*(1), 7201195030p7201195031-7201195030p7201195037.

Martin, N. A. (2010). *Test of Visual Motor Skills - 3rd Edition*. Western Pyschological Services.

Martin, N. A. (2017). *Test of Visual Perceptual Skills - 4th Edition*. ATP Assessments.

May-Benson, T. A. (2005). *Examining ideational abilities in children with dyspraxia*. Boston University. Proquest Dissertations. https://ezproxy2.library.colostate.edu/login?url=https://www.proquest.com/dissertations-theses/examining-ideational-abilities-children-with/docview/305027609/se-2

May-Benson, T. A., & Koomar, J. A. (2010). Systematic review of the research evidence examining the effectiveness of interventions using a sensory integrative approach for children. *American Journal of Occupational Therapy*, *64*(3), 403-414.

May-Benson, T. A., & Teasdale, A. (2021). Inter-Rater and Test-Retest Reliability of the Sensory Integration Clinical Observations. *Physical & occupational therapy in pediatrics*, *41*(1), 74-84.

Meijer R.R., & Sitsma K. (2001). *Person-fit statistic—what is their purpose*. https://www.rasch.org/rmt/rmt152d.htm

Menier, C., Forget, R., & Lambert, J. (1996). Evaluation of two-point discrimination in children: reliability, effects of passive displacement and voluntary movements.Developmental Medicine & Child Neurology, 38(6), 523-537. https://doi.org/10.1111/j.1469-8749.1996.tb12113.x

Miller, L. J. (2006). *Miller function & participation scales*. Pearson.

Miller, L. J., Anzalone, M. E., Lane, S. J., Cermak, S. A., & Osten, E. T. (2007). Concept evolution in sensory integration: a proposed nosology for diagnosis. *American Journal of Occupational Therapy*, 61(2), 135-140. https://doi.org/10.5014/ajot.61.2.135

Mohammadi, F., Bahram, A., Khalaji, H., Ulrich, D. A., & Ghadiri, F. (2019). Evaluation of the psychometric properties of the Persian version of the Test of Gross Motor Development–3rd edition. *Journal of Motor Learning and Development*, *7*(1), 106-121.

Mokkink, L. B., Prinsen, C., Patrick, D. L., Alonso, J., Bouter, L., de Vet, H. C., & Terwee, C.B. (2018). *COSMIN methodology for systematic reviews of patient-reported outcome measures (PROMs)*. User manual, 78(1).

Mulligan, S. (1998). Patterns of Sensory Integration Dysfunction: A Confirmatory Factor Analysis. *The American Journal of Occupational Therapy*, 52(10), 819-828. https://doi.org/https://doi.org/10.5014/ajot.52.10.819

Mulligan, S. (2011). Validity of the Postrotary Nystagmus Test for Measuring Vestibular Function. *OTJR: Occupation, Participation and Health*, *31*, 97-104. https://doi.org/10.3928/15394492-20100823-02

Mulligan, S. (2020). Assessment of Sensory Integration Functions Using the Sensory Integration and Praxis Tests. In A. C. Bundy & S. J. Lane (Eds.), *Sensory Integration: Theory and Practice* (3rd ed.). F. A. Davis.

Murray, E. A., Cermak, S. A., & O'Brien, V. (1990). The relationship between form and space perception, constructional abilities, and clumsiness in children. *The American Journal of Occupational Therapy*, 44(7), 623-628. https://doi.org/10.5014/ajot.44.7.623

Mutti, M., Martin, N., Sterling, H., Spalding, N. (2017). *Quick Neurological Screening Test, 3rd Edition, Revised*. ATP Assessments.

Ottenbacher, K. (1982). Sensory integration therapy: Affect or effect. *American Journal of Occupational Therapy*, *36*(9), 571-578.

Parham, L. D., & Cosbey, J. (2020). Sensory Integration in Everyday Life. In A. C. Bundy & S. J. Lane (Eds.), *Sensory Integration: Theory and Practice* (3rd ed.). F. A. Davis.

Parham, L. D., Ecker, C., Kuhaneck, H., Henry, D. A., & Glennon, T. J. (2007). *Sensory Processing Measure*. Western Psychological Services.

Parham, L. D., Roley, S. S., May-Benson, T. A., Koomar, J., Brett-Green, B., Burke, J. P., Cohn, E. S., Mailloux, Z., Miller, L. J., & Schaaf, R. C. (2011). Development of a fidelity measure for research on the effectiveness of the Ayres Sensory Integration® intervention. *American Journal of Occupational Therapy*, *65*(2), 133-142.

Parks, K., Schulz, S., McDonnell, C. G., Anagnostou, E., Nicolson, R., Kelley, E., Georgiades, S., Crosbie, J., Schachar, R., Liu, X., & Stevenson, R. (2020). Sensory Processing in ASD and ADHD: A Confirmatory Factor Analysis.https://doi.org/https://doi.org/10.31234/osf.io/myjbq

Pfeiffer, B., Moskowitz, B., Paoletti, A., Brusilovskiy, E., Zylstra, S. E., & Murray, T. (2015). Developmental Test of Visual–Motor Integration (VMI): an effective outcome measure for handwriting interventions for kindergarten, first-grade, and second-grade students? *American Journal of Occupational Therapy*, *69*(4), 6904350010p6904350011-6904350010p6904350017.

Polatajko, H. J., Kaplan, B. J., & Wilson, B. N. (1992). Sensory integration treatment for children with learning disabilities: Its status 20 years later. *The Occupational Therapy Journal of Research*, *12*(6), 323-341.

Portney, L. G. (2020). *Foundations of clinical research: applications to evidence-based practice*. FA Davis.

Prinsen, C. A., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., De Vet, H. C., & Terwee, C. B. (2018). COSMIN guideline for systematic reviews of patient-reported outcome measures. *Quality of Life Research*, *27*(5), 1147-1157.

Psotta, R., & Abdollahipour, R. (2017). Factorial Validity of the Movement Assessment Battery for Children—2nd Edition (MABC-2) in 7-16-Year-Olds. *Perceptual and motor skills*, *124*(6), 1051-1068.

Rais, M., Binder, D. K., Razak, K. A., & Ethell, I. M. (2018). Sensory Processing Phenotypes in Fragile X Syndrome. *ASN Neuro*, *10*, 1759091418801092. https://doi.org/10.1177/1759091418801092

Rihtman, T., & Parush, S. (2014). Suitability of the Miller Function and Participation Scales (M–FUN) for use with Israeli children. *American Journal of Occupational Therapy*, *68*(1), e1-e12.

Rintala, P. O., Sääkslahti, A. K., & Iivonen, S. (2017). Reliability assessment of scores from video-recorded TGMD-3 performances. *Journal of Motor Learning and Development*, *5*(1), 59-68.

Saraiva, L., Rodrigues, L. P., Cordovil, R., & Barreiros, J. (2013). Motor profile of Portuguese preschool children on the Peabody Developmental Motor Scales-2: A cross-cultural study. *Research in Developmental Disabilities*, *34*(6), 1966-1973.

Schaaf, R. C., Dumont, R. L., Arbesman, M., & May-Benson, T. A. (2018). Efficacy of occupational therapy using Ayres Sensory Integration®: A systematic review. *American Journal of Occupational Therapy*, *72*(1), 7201190010p7201190011-7201190010p7201190010.

Schaaf, R. C., & Lane, A. E. (2015). Toward a Best-Practice Protocol for Assessment of Sensory Features in ASD. Journal of Autism and Developmental Disorders, 45(5), 1380-1395. https://doi.org/10.1007/s10803-014-2299-z

Schaaf, R. C., & Mailloux, Z. (2015). Clinician's guide for implementing Ayres Sensory Integration: promoting participation for children with autism. AOTA Press, The American Occupational Therapy Association, Inc.

Schoemaker, M. M., Niemeijer, A. S., Flapper, B. C., & Smits-Engelsman, B. C. (2012). Validity and reliability of the movement assessment battery for children-2 checklist for children with and without motor impairments. *Developmental Medicine & Child Neurology*, *54*(4), 368-375.

Schubert, M. C., & Zee, D. S. (2010). Saccade and vestibular ocular motor adaptation. *Restor Neurol Neurosci*, *28*(1), 9-18. https://doi.org/10.3233/RNN-2010-0523

Schulz, J., Henderson, S. E., Sugden, D. A., & Barnett, A. L. (2011). Structural validity of the Movement ABC-2 test: Factor structure comparisons across three age groups. *Research in Developmental Disabilities*, *32*(4), 1361-1369.

Serbetar, I., Loftesnes, J. M., & Mamen, A. (2019). Reliability and Structural Validity of the Movement Assessment Battery for Children-2 in Croatian Preschool Children. *Sports*, *7*(12), 248.

Simons, J., & Eyitayo, G. (2016). Aspects of reliability and validity of the TGMD-3 in 7-10 year old children with intellectual disability in Belgium. *European Psychomotricity Journal*, *8*(1), 3-16.

Skard, G., & Bundy, A. C. (2008). Test of Playfulness In P. D. L. & F. L. S. (Eds.), *Play in occupational therapy for children* (pp. 71–93). Mosby Elsevier.

Smith, R. M., & Plackner, C. (2009). The family approach to assessing fit in Rasch measurement. *J Appl Meas*, *10*(4), 424-437. https://www.ncbi.nlm.nih.gov/pubmed/19934529

Smith Roley, S., Mailloux, Z., & Erwin, B. (2022). *Ayres sensory integration*. Sensory Integration Global Network (SIGN). Retrieved from https://www.siglobalnetwork.org/ayres-sensory-integration.

Smits-Engelsman, B. C., Niemeijer, A. S., & van Waelvelde, H. (2011). Is the Movement Assessment Battery for Children-a reliable instrument to measure motor performance in 3 year old children? *Research in Developmental Disabilities*, *32*(4), 1370-1377.

Spasić M., B. M., Lukač J. (2022). Differences in Balance with Eyes Closed, Eyes Opened and Virtual Reality Environment: A pilot-study. *Journal of Anthropology of Sport and Physical Education*, *6*(3), 11-14. https://doi.org/10.26773/jaspe.220702

Szklut, S. (2010). Using clinical reasoning to evaluate sensory processing dysfunction. *AOTA: Sensory Integration Special Interest Section*, *33*(4), 1-4.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International journal of medical education*, *2*, 53.

Temple, V. A., & Foley, J. T. (2017). A peek at the developmental validity of the Test of Gross Motor Development–3. *Journal of Motor Learning and Development*, *5*(1), 5-14.

Ulrich, D. (2019). TGMD-3: test of gross motor Development—third edition. In: Pro-Ed Austin, TX.

Valentini, N., Ramalho, M., & Oliveira, M. (2014). Movement Assessment Battery for Children-2: Translation, reliability, and validity for Brazilian children. *Research in Developmental Disabilities*, *35*(3), 733-740.

Valentini, N. C., & Zanella, L. W. (2022). Peabody Developmental Motor Scales-2: The Use of Rasch Analysis to Examine the Model Unidimensionality, Motor Function, and Item Difficulty. *Front Pediatr*, *10*, 852732. https://doi.org/10.3389/fped.2022.852732

Valentini, N. C., Zanella, L. W., & Webster, E. K. (2017). Test of Gross Motor Development—Third edition: Establishing content and construct validity for Brazilian children. *Journal of Motor Learning and Development*, *5*(1), 15-28.

van Jaarsveld, A., Mailloux, Z., & Herzberg, D. S. (2012). The use of the Sensory Integration and Praxis Tests with South African children. *South African Journal of Occupational Therapy*, *42*(3), 12-18. http://www.scielo.org.za/scielo.php?script=sci_arttext&pid=S2310-38332012000300004

Van Waelvelde, H., Peersman, W., Lenoir, M., & Engelsman, B. C. S. (2007). Convergent validity between two motor tests: movement-ABC and PDMS-2. *Adapted Physical Activity Quarterly*, *24*(1), 59-69.

Vanvuchelen, M., Roeyers, H., & De Weerdt, W. (2011). Development and initial validation of the Preschool Imitation and Praxis Scale (PIPS). *Research in Autism Spectrum Disorders*, *5*(1), 463-473.

Vanvuchelen, M., Roeyers, H., & De Weerdt, W. (2011). Objectivity and stability of the Preschool Imitation and Praxis Scale. *American Journal of Occupational Therapy*, *65*(5), 569-577.

Vanvuchelen, M., & Vochten, C. (2011). How much change is true change? The smallest detectable difference of the Preschool Imitation and Praxis Scale (PIPS) in preschoolers with intellectual disabilities of heterogeneous aetiology. Res Dev Disabil, 32(1), 180-187. https://doi.org/10.1016/j.ridd.2010.09.019

Vargas, S., & Camilli, G. (1999). A meta-analysis of research on sensory integration treatment. *American Journal of Occupational Therapy*, *53*(2), 189-198.

Wagner, M. O., Webster, E. K., & Ulrich, D. A. (2017). Psychometric properties of the Test of Gross Motor Development, (German translation): Results of a pilot study. *Journal of Motor Learning and Development*, *5*(1), 29-44.

Webster, E. K., & Ulrich, D. A. (2017). Evaluation of the psychometric properties of the Test of Gross Motor Development—third edition. *Journal of Motor Learning and Development*, *5*(1), 45-58.

Wilson, B. N. (2004). *Clinical Observations of Motor and Postural Skills (COMPS)*. École des sciences de la réadaptation, Sciences de la santé, Université d'Ottawa.

Wilson, B. N., Pollock, N., Kaplan, B., & Law, M. (2000). The Clinical Observation of Motor and Postural Skills – Second Edition (COMPS-2). Therapro, Inc.

Wright, B. D. (1991). Diagnosing Misfit. https://www.rasch.org/rmt/rmt52k.htm

Wright, B. D. (1998). *Estimating Rasch Measures for Extreme Scores*. https://www.rasch.org/rmt/rmt122h.htm

Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis*. MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best Test Design*. MESA Press.

Wuang, Y.-P., Lin, Y.-H., & Su, C.-Y. (2009). Rasch analysis of the Bruininks–Oseretsky Test of Motor Proficiency-in intellectual disabilities. *Research in Developmental Disabilities*, *30*(6), 1132-1144.

Wuang, Y. P., & Su, C. Y. (2009). Reliability and responsiveness of the Bruininks–Oseretsky Test of Motor Proficiency-in children with intellectual disability. *Research in Developmental Disabilities*, *30*(5), 847-855.

Wuang, Y. P., Su, C. Y., & Huang, M. H. (2012). Psychometric comparisons of three measures for assessing motor functions in preschoolers with intellectual disabilities. *Journal of Intellectual Disability Research*, *56*(6), 567-578.

Wuang, Y. P., Su, J. H., & Su, C. Y. (2012). Reliability and responsiveness of the Movement Assessment Battery for Children–Second Edition Test in children with developmental coordination disorder. *Developmental Medicine & Child Neurology*, *54*(2), 160-165.

Zingaretti, P., Petta, A. M., Cruciani, G., & Spitoni, G. F. (2019). Tactile sensitivity, tactile acuity, and affective touch: from childhood to early adolescence. Somatosensory & Motor Research, 36(1), 90-96. https://doi.org/10.1080/08990220.2019.1604334

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement*, *36*(1), 1-2.

# APPENDIX A: OCCUPATIONAL AND REHABILITATION SCIENCE

*You will graduate with a Doctor of Philosophy in Occupation and Rehabilitation Science (ORS), and must be able to, therefore, situate your research within the context of ORS. To this end, how might your anticipated dissertation research specifically draw from existing literature and conceptual perspectives of occupation science and rehabilitation science? What contributions do you believe your anticipated dissertation research will make to these fields and why are those contributions needed/important?*

Scientific disciplines define valid areas of inquiry, tools, and approaches for scholars. Each discipline has, at its core, a topic, concept, or idea that drives researchers. In this chapter, I situated my dissertation in two complementary but distinct disciplines: occupational science and rehabilitation science. Each discipline has a unique grounding concept, history, and set of methodologies. In the sections that follow, I defined occupational science and rehabilitation science, drawing upon seminal literature. I explained how I situate my study within each field. I described the influences of each discipline on my anticipated dissertation and the contributions my research will make.

## The EASI: My Anticipated Dissertation

The Evaluation in Ayres Sensory Integration® (EASI; Mailloux et al., 2018) is a suite of 20 performance-based tests designed to measure four constructs of sensory integration: sensory perception, sensory reactivity, praxis, and postural/ocular/bilateral integration. The EASI is conceptually aligned with a theory and approach to occupational therapy practiced known as sensory integration. Currently, normative data collection is underway in more than 80 different countries. For my anticipated dissertation research, I will evaluate the validity and reliability of this normative data, using a psychometric model called the Rasch model. Validating this large

dataset will present an invaluable contribution to two scientific disciplines: occupational science and rehabilitation science.

## Occupational Science

Yerxa and colleagues (1989) first conceptualized occupational science as the study of the human as an occupational being. They proposed occupational science as a basic scientific discipline essential to support the application of occupational therapy. More modern conceptualizations have shifted the focus away from individual humans, centralizing occupation itself as a transactional, multidimensional human enterprise (Clark et al., 1991; Cutchin & Dickie, 2012). As the discipline has grown, scholars have rejected the dualism of basic versus applied, arguing instead that valid occupational science studies may fall along a gradient of these two concepts (Molineux & Whiteford, 2011). A modern definition of occupational science, presented by Rudman and colleagues (2010), reflected these new conceptualizations: "a vibrant academic discipline that will engage in a continuum of knowledge generation and action concerning the construct of occupation" (p. 2008).

Zemke and Clark (1996) described relevant areas of knowledge germane to the discipline of occupational science: the *substrates*, *form*, *function*, and *meaning* of occupation. *Substrates* refer to the human capacities required for engagement in occupation. *Form* refers to observable features of an occupation (e.g., the sequence of steps or materials required for a particular occupation). *Function* refers to the outcomes of participating in an occupation (e.g., increasing self-esteem, enabling independence, strengthening relationships). *Meaning* refers to the significance of occupation, including societal, personal, cultural, and historical expressions. The range of knowledge in occupational science lends itself to epistemological diversity in occupational science research. Studies of the substrates and form of occupation may be best

served by postpositivist (often quantitative) approaches (e.g., Hernandez et al., 2020; Wood, 2005). Studies of the function and meaning of occupation in individuals, families, and communities are often interpretivist (qualitative) in nature (e.g., Shank & Cutchin, 2010).

## Situating my Anticipated Dissertation in Occupational Science

My anticipated dissertation is best situated among the *substrates* of occupation. Sensory integration is a key physiological process that facilitates occupational performance and participation (Parham & Cosbey, 2020).

Occupation is the central construct that unites occupational science. Although my psychometrics studies will not address occupational directly, occupation *is* at the center of my inquiry. My core motivation for developing the EASI is to create an instrument that can enable occupational scientists and therapists to better understand and facilitate children's performance and participation in daily, valued occupations.

## Influences from Occupational Science

Two topics from the discipline of occupational science are particularly relevant to my dissertation research: (1) sensory integration as a substrate of occupation, and (2) occupational justice.

### Sensory Integration as a Substrate of Occupation

Yerxa and colleagues (1989) acknowledged sensory integration theory as an early example of occupational science. This theory began with Ayres in the mid 20th century. She defined sensory integration as "the organization of sensory input for use" (Ayres, 1979). Per Ayres' theory, sensory integration is a key physiological process that enables people to engage with their environments. Between 1965 and 1989, Ayres performed a series of factor analytic studies using a suite of instrument she developed called the Sensory Integration and Praxis Tests

(SIPT; Lane, Bundy & Gorman, 2020). Using these studies, Ayres and her colleagues developed, refined, and modified a theory of sensory integration that connected neurological processes to observable behavior in children. Ayres' colleagues and successors continued her research through both basic and applied studies that contribute to the theory, and, therefore, to the body of knowledge in occupational science (e.g., Mailloux et al., 2011; Mulligan, 1998).

My dissertation work will draw from the tradition of research that has emerged from Ayres' work. I will draw upon this research as I evaluate the psychometric properties of the EASI by making decisions to remove or retain items based on their role in the evolving theoretical constructs of sensory integration. Specifically, I aim to ensure that all constructs of modern sensory integration theory, as generated by Ayres and colleagues and as summarized by Bundy & Lane (2020b), are adequately and equitably represented in the instrument.

**Occupational Justice**

Occupational justice is an evolving concept in occupational science (Hocking, 2017). Broadly, occupational justice refers to three core beliefs: (1) that all humans are innately occupational, (2) that occupation is influenced by societal and structural factors (e.g., policy, cultural values), and (3) that occupational can improve the circumstances of vulnerable people (Hocking, 2017). Given these core beliefs, occupational justice demands equitable opportunity for occupation for all humans regardless of age, national origin, disability status, or past behavior (Stadnyk et al., 2010).

My dissertation reflects my alignment with the concept of occupational justice. Disparities in healthcare are a source of occupational injustice – especially for children with disabilities (Bass-Haugen, 2009). For children who live in families who can access and afford healthcare, sensory integration assessment and treatment may be readily available. This

treatment may reduce occupational inequity by providing the child with the ability to produce adaptive responses that enable occupational performance and participation. However, children in underserved communities may not have ready access to treatment – and the treatment they do receive may come from underfunded clinics that lack the resources to properly evaluate sensory integration deficits. The EASI test kit, unlike previous performance-based measures of sensory integration, is comprised of affordable, accessible materials. The training can be completed online. The instrument is available in 15 languages. Therefore, the EASI may contribute to closing the healthcare gaps for children with sensory integration challenges. My dissertation will provide key research in the creation of an assessment that contributes to a vision of occupational justice for all children for whom sensory integration is a concern.

## Contributions to Occupational Science

My anticipated dissertation is an important step to making the EASI an available test for use in pediatric occupational therapy for children with sensory integration concerns. The widespread use of the EASI has the potential to contribute to the body of occupational science knowledge in three main ways: (1) by creating a stronger theory of sensory integration; (2) by translating theoretical knowledge into occupational therapy practice; and (3) by enabling a better understanding of childhood occupations.

### A Stronger Theory of Sensory Integration

My dissertation will provide a valuable contribution to the body of occupational science literature. Sensory integration is a crucial biological substrate for occupation. The EASI will contribute to sensory integration theory by producing valid, reliable data to evaluate and continue to revise this theory. Currently, most theory development in sensory integration uses the SIPT (Mulligan, 2020). Unlike the SIPT, the EASI will have normative data collected from more than

80 countries and in 15 languages. Therefore, the EASI will enrich theory development by allowing scholars to consider cultural influences as well as universal patterns. This is a relevant, timely endeavor as we seek to make knowledge in occupational science more culturally-relevant and globally-useful (Frank, 2010). My response to question 2 describes additional advantages of the EASI instrument for sensory integration theory development.

**Translation into Occupational Therapy Practice**

In addition to contributing to theory, my research will contribute to knowledge translation into the practice of occupational therapy. Kielhofner (2005) argued that the scholars within the discipline of occupational science must ensure that their research enables translation into practice, most often in the field of occupational therapy. Sensory integration is among the most common frames of reference for pediatric occupational therapists (Schaaf et al., 2018). The practice of sensory integration requires thorough assessment of children's sensory function (Parham et al., 2011). The EASI will be an open-source, inexpensive assessment available to occupational therapy practitioners all over the world. By examining the validity and reliability of data collected with this instrument, I aim to enable practitioners to bring this tool into their practice.

**A Better Understanding of Childhood Occupation**

Childhood occupations such as play, education, rest/sleep, and activities of daily living rely upon sensory integration (Parham & Cosbey, 2020). Although my dissertation study does not directly address these occupations, my work will make a valuable contribution to the body of knowledge investigating childhood occupations. The EASI, used alongside participation or functional measures, can be a source of knowledge about the nature of occupational function/dysfunction and participation for children with sensory integration dysfunction.

## Rehabilitation Science

In addition to occupational science, my anticipated dissertation both draws upon and will contribute to the discipline of rehabilitation science. Brandt and Pope, authors of the Institute of Medicine's (1997) *Enabling America*, described rehabilitation science as "the study of movement among states in the enabling-disabling process" (p. 25). This process encompasses four "states" – pathology, impairment, functional limitation, and disability – drawn from the work of Nagi (1976). *Pathology* studies examine cellular and tissue abnormalities and processes. *Impairment* studies examine organs and organ systems with special consideration to atypical states that result in loss of function. *Functional limitation* studies examine the impacts of physiological dysfunction on human activity. *Disability* studies examine the intersections of these functional limitations with lived contexts. The *disabling* process occurs when pathology and impairment give rise to functional limitations within a person's context (i.e., disability). *Enabling* occurs when a person with a pathology and/or impairment can retain function because of a supported environment and/or effective treatment. Similar to occupational science, rehabilitation science is both basic and applied. Basic studies examine the states described above, while applied studies draw upon this knowledge to create and evaluate rehabilitation techniques. Given the breadth of these constructs, rehabilitation science naturally draws upon other disciplines of study. Studies from diverse fields, ranging from neuroscience to implementation science, from cellular biology to sociology, can be included among rehabilitation science literature, as long as they generate knowledge related to the enabling-disabling process.

Like occupational science, rehabilitation science lends itself to a range of epistemological approaches. Studies regarding pathology and impairment are mainly postpositivist and quantitative in nature (e.g., Crasta et al., 2018), while disability studies are often suited for the

interpretivist and participatory action domains (generally using qualitative methods; e.g., Schneider, 2012). Functional limitation research takes many epistemological approaches, and mixed methods are sometimes appropriate, especially when examining rehabilitation interventions (e.g., Pergolotti et al., 2012).

**Situating my Anticipated Dissertation in Rehabilitation Science**

Rehabilitation science also provides a conceptual groundwork and motivation for my research. I situate my anticipated dissertation within the *impairment* level of Brandt and Pope's (1997) four levels of rehabilitation science. Sensory integration concerns represent a primary dysfunction in the nervous system that can result in functional limitations (Bundy & Lane, 2020b). Developing a measure of sensory integration, is, therefore, a valid endeavor for rehabilitation science.

Just as occupation is central to occupational science, the enabling-disabling process lies at the heart of rehabilitation science. By evaluating the EASI, I aim to create an instrument that facilitates the development of evidence-based rehabilitation techniques. Sensory integration therapy (based on a theoretically- and psychometrically-sound assessment tool – the EASI) can move children towards the enabled end of the enabling-disabling process by addressing fundamental impairments that give rise to functional limitations and disability. As a rehabilitation science study, my dissertation will both draw from and contribute to this discipline.

**Influences from Rehabilitation Science**

Two primary bodies of knowledge in rehabilitation science will influence my anticipated dissertation: (1) the neuroscience of sensory integration; and (2) best practices in assessment for pediatric rehabilitation.

### Neuroscience of Sensory Integration

In recent years, neuroscientists in the rehabilitation field have begun to explore neural processes in relation to sensory integration (e.g., Lane et al., 2019; Kilroy et al., 2019). These studies have confirmed many of Ayres' hypotheses and challenged or built upon on others. Reviews of the neuroscience literature suggest that the processes underlying sensory integration are highly complex and involve many brain structures and networks – this creates a challenge in relating clinical outcomes of sensory integration therapy with neural changes (Kilroy et al., 2019). It will be crucial for me to be grounded in these neural underpinnings as I examine the EASI. For example, if I find that a test of balance has many misfitting items (i.e., demonstrates excessive "noise"), I must consider the possibility that neural processes such as attention, arousal, and vestibular/proprioceptive systems are intersecting in variable ways across participants. Knowledge of basic neuroscience is essential to the study of sensory integration.

### Assessment in Pediatric Rehabilitation

Additionally, literature exploring clinicians' use (or lack thereof) of assessments and outcome measures in pediatric rehabilitation influences my anticipated dissertation (e.g., Auld & Johnston, 2018; Hanna et al., 2011; King et al., 2007; Law, 2003). Taken together, this body of literature suggests that pediatric assessment tools must (1) be efficient, (2) provide meaningful information for clinicians, (3) foster conversation and relationship with clients and families, and (4) demonstrate strong psychometric properties. As I recommend revisions to strengthen the EASI, I will keep each of these principles at the forefront of my efforts. In the spirit of efficiency, I will suggest the removal of redundant items. At the same time, I will work to ensure that the retained items create a complete, meaningful picture of children's abilities that fosters clinician-client engagement and shared goal setting.

## Contributions to Rehabilitation Science

My dissertation will also contribute knowledge to at least two fields of study in rehabilitation science: (1) sensory integration as a part of the enabling-disabling process; and (2) development and evaluation of rehabilitation practices for children with sensory integration challenges.

### Sensory Integration Along the Enabling-Disabling Process

The EASI will contribute to knowledge along all four states of the enabling-disabling process (pathology, impairment, functional limitations, and disability). The EASI will provide a valid and reliable way for rehabilitation scientists to examine the observable behaviors associated with sensory integration impairments. Sophisticated neuroimaging techniques may be paired with the EASI to connect observable sensory features with pathology. In the other direction, the EASI may be used in conjunction with time-use measures or qualitative methods to better understand the impact of sensory integrative dysfunction on functional limitations and disability.

### Development and Evaluation of Rehabilitation Techniques for Sensory Integration

As Bundy and Lane (2020a) confirmed, the efficacy of sensory integration therapy remains somewhat nebulous. While the authors advocate for the use of occupational outcomes to evaluate sensory integrative therapy (e.g., goal attainment scaling), impairment-level outcome measures such as the EASI are also essential in understanding the *mechanisms* by which sensory integration can facilitate occupational performance. Determining validity and reliability is a key piece to creating a valid outcome measure. My study alone will not demonstrate the acceptability of the EASI as an outcome measure – this will require studies examining change over time and with intervention. However, my study will lay the groundwork for studies that can determine the

responsiveness of the EASI, and in turn, the efficacy of sensory integration as an intervention approach.

## Conclusion

As a scholar of occupational science *and* rehabilitation science, I am pursuing a multidisciplinary research degree. Although I benefit from the diversity of approaches and viewpoints inherent in multidisciplinary research, I am also tasked with the responsibility of maintaining *both* disciplines at the core of my anticipated dissertation. In this response, I described the influence of both occupational science and rehabilitation science on my scholarly work. As I proceed into the next phase of my research, both occupation and the enabling-disabling process will inform my process.

# REFERENCES


Auld, M. L., & Johnston, L. M. (2018). A touchy topic: tactile assessment among pediatric therapists. *Disability and Rehabilitation, 40*(3), 267-276.

Bass-Haugen, J. D. (2009). Health dispartieis: Examination of evidence relevant for occupational therapy. *American Journal of Occupational Therapy, 63*, 24-34.

Brandt, E. N., & Pope, A. M. (1997). *Enabling America: Assessing the role of rehabilitation science and engineering.* Washington, DC: The National Academies Press.

Breckler, S. (2005, October). *The importance of disciplines*. Retrieved 2020 August, from American Pyschological Association: https://www.apa.org/science/about/psa/2005/10/ed-column

Bundy, A. C., & Lane, S. J. (2020). Is sensory integration effective? A complicated question to end the book. In A. C. Bundy, & S. J. Lane (Eds.), *Sensory Integration: Theory and Practice* (3rd ed., pp. 568-577). Philadelphia, PA: FA Davis.

Bundy, A. C., & Lane, S. J. (2020). Sensory integration: A. Jean Ayres' theory revisited. In A. C. Bundy, & S. J. Lane (Eds.), *Sensory Integration: Theory and Practice* (3rd ed., pp. 1-20). Philadelphia, PA: FA Davis.

Clark, F., & Lawlor, M. C. (2009). The making and mattering of occupational science. In E. B. Crepeau, E. S. Cohn, & B. A. Schell (Eds.), *Willard & Spackman's Occupational Therapy* (11th ed., pp. 2-14). Philadelphia, PA: Lippincott William & Wilkins.

Clark, F., Zemke, R., Frank, G., Parham, D., Neville-Jan, A., Hedricks, C., . . . Abreu, B. (1993). The issue is - Dangers inherent in the partition of occupational therapy and occupational science. *American Journal of Occupational Therapy, 47*(2), 184-186.

Crasta, J., Thaut, M. H., Anderson, C. W., Davies, P. L., & Gavin, W. J. (2018). Auditory priing improves neural synchoronization in auditory-motor entrainment. *Neuropsychologia, 117*, 102-112.

Cutchin, M. P., & Dickie, V. A. (2012). Transactionalism: Occupational science and the pragmatic attitude. In G. E. Whiteford, & C. Hocking (Eds.), *Occupational science: Society, iclusion, participation* . Oxford, UK: Blackwell Publishing.

Frank, G. (2010). The 2010 Ruth Zemke lecture in occupational science: Occupational therapy/occupational science/occupational justice: Moral commitments and global assemblages. *Journal of Occupational Science, 19*(1), 25-35.

Hanna, S. E., Russell, D. J., Bartlet, D. J., Kertoy, M. L., Rosenbaum, P. L., & Wynn, K. (2007). Measurement practices in pediatric rehabilitation: A survey of physical therapists, occupational therapists, and speech-language pathologists in Ontario. *Phsyical & Occupational Therapy in Pediatrics, 27*(2), 25-42.

Hernandez, R., Vidmar, A., & Pyatak, E. A. (2020). Lifestyle balance, restful and strenuous occupations, and physiological activation. *Journal of Occupational Science*, 1-16.

Hocking, C. (2017). Occupational justice as social justice: The moral claim for inclusion. *Journal of Occupational Science, 24*(1), 29-42.

Kielhofner, G. (2005). Scholarship and practice: Bridging the divide. *American Journal of Occupational Therapy, 59*(2), 231-239.

Kilroy, E., Aziz-Zadeh, L., & Cermak, S. (2019). Ayres theories of autism and sensory integration revisited: What contemporary neuroscience has to say. *Brain Sciences, 9*(68), 1-20.

King, G., Wright, V., & Russell, D. J. (2011). Understanding pediatric rehabilitation therapists' lack of use of outcome measures. *Disability and Rehabilitation, 33*, 25-26.

Lane, S. J., Bundy, A. C., & Gorman, M. E. (2020). Composing a theory. In *Sensory Integration: Theory and Practice* (pp. 40-55). Philadelphia, PA: FA Davis.

Lane, S. J., Mailloux, Z., Schoen, S., Bundy, A., May-Benson, T. S., Parham, L. D., . . . Schaaf, R. C. (2019). Neural foundations of Ayres Sensory Integration. *Brain Science, 9*(7), 1-14.

Law, M. (2003). Outcome measurement in pediatric rehabilitation. *Physical & Occupational Therapy in Pediatrics, 23*(3), 1-4.

Mailloux, Z., Mulligan, S., Smith Roley, S., Blance, E., Cermak, S., Coleman, G. C., . . . Lane, C. J. (2011). Verification and clarification of patterns of sensory integrative dysfunction. *American Journal of Occupational Therapy, 65*(2), 143-151.

Mailloux, Z., Parham, L. D., Smith Roley, S., Ruzzano, L., & Schaaf, R. C. (2018). Introduction to the Evaluation in Ayres Sensory Integration (EASI). *American Journal of Occupational Therapy, 72*, 1-7.

Molineux, M., & Whiteford, G. (2011). Occupational science. In E. A. Duncan (Ed.), *Foundations for Practice in Occupational Therapy* (5th ed., pp. 261-273). New York, NY: Elsevier.

Mulligan, S. (1998). Patterns of sensory integration dysfunction: A confirmatory factor analysis. *American Journal of Occupational Therapy, 52*(10), 819-828.

Mulligan, S. (2020). Assessment of sensory integration functions using the Sensory Integration and Praxis Tests. In *Sensory Integration: Theory and Practice* (3rd ed., pp. 208-221). Philadelphia, PA: FA Davis.

Nagi, S. Z. (1976). An epidemiology of disability among adults in the United States. *The Milbank Memorial Fund Quarterly: Health and Society, 54*(4), 439-467.

Parham, L. D., & Cosbey, J. (2020). Sensory integration in everyday life. In A. C. Bundy, & S. J. Lane (Eds.), *Sensory Integration: Theory and Practice* (3rd ed., pp. 21-39). Philadelphia, PA: FA Davis.

Parham, L. D., Smith Roley, S., May-Benson, T. A., Koomar, J., Brett-Green, B., Burke, J. P., . . . Schaaf, R. C. (2011). Development of a fidelity measure of research on the effectiveness of the Ayres Sensory Integration intervention. *American Journal of Occupational Therapy, 65*(2), 133-142.

Pergolotti, M., Bailliard, A., McCarthy, L., Farley, E., Covington, K. R., & Doll, K. M. (2020). Women's experiences after ovarian cancer surgery: Distress, uncertainty and the need for occupational therapy. *American Journal of Occupational Therapy, 73*(4), 1-9.

Schaaf, R. C., Dumont, R. L., Arbesman, M., & May-Benson, T. A. (2018). Efficacy of occupational therapy using Ayres Sensory Integration: A systematic review. *American Journal of Occupational Therapy, 72*(1), 1-10.

Schneider, B. (2012). Participation action research, mental health service user research, and the hearing (our) voices projects. *International Journal of Qualitative Methods, 11*(2), 152-165.

Shank, K. H. (2013). Mixed methods and pragmatism for research on occupation. In *Transactional Perpsectives on Occupation* (pp. 183-195). New York: Springer.

Shank, K. H., & Cutchin, M. P. (2010). Transactional occupations of older women aging-in-pace: Negotiating change and meaning. *Journal of Occupational Science, 17*(1), 4-13.

Stadnyk, R., Townsend, E., & Wilcock, A. (2010). Occupational justice. In C. H. Christiansen, & E. A. Townsend (Eds.), *Introduction to occupation: The art and science of living* (pp. 329-358). Upper Saddle River, NJ: Pearson Education.

Wood, W. (2005). Toward developing new occupational science measures: An example from dementia care research. *Journal of Occupational Science, 12*(3), 121-129.

Wright-St Clair, V. A., & Hocking, C. (2014). In B. A. Schell, G. Gillen, M. E. Scaffa, & E. S. Cohn (Eds.), *Willard & Spackman's Occupational Therapy* (12th ed., Vols. 82-93). Philadelphia, PA: Lippincot William &Wilkins.

Yerxa, E. J., Clark, F., Frank, G., Jackson, J., Parham, L. D., Pierce, D., . . . Zemke, R. (1989). An introduction to occupational science, a foundation for occupational therapy in the 21st century. *Occupational Therapy in Health Care, 6*(4), 1-17.

Zemke, R., & Clark, F. (1996). Preface. In R. Zemke, & F. Clark (Eds.), *Occupational science: The evolving discipline* (pp. vii-xviii). Philadelphia, PA: FA Davis.

APPENDIX B

The tables that follow contain each of the studies I examined for this review of literature. In these tables, "mq" refers to methodological quality, as rated by the COSMIN framework (VG = very good, A = adequate, D = doubtful, I = inadequate). "n" refers to the number of participants, and "r" (short for "rating") refers to the adequacy of the evidence using COSMIN standards (+, -, +/-, or ?; see *Methodological Framework*). Samples comprise typically developing children unless otherwise noted. **Bold** studies are drawn from the instrument manuals. All year groups are inclusive unless otherwise noted (i.e., 3-6 years includes children 3 years, 0 months to 6 years, 11 months). If instruments from the review are not included in these tables, I did not find studies supporting any of their measurement properties.

**Table 1**

*Studies Evaluating the Reliability of SIPT*

| SIPT Test | Study | Country | Sample | Test-retest Reliability | | | Inter-rater Reliability | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | mq | n | r | mq | n | r |
| SV | **Ayres (2005)** | US | 4-8 years, TD and SID | A | 49 | - | A | 63 | + |
| FG | **Ayres (2005)** | US | 4-8 years, TD and SID | A | 47 | - | A | 58 | + |
| MFP | **Ayres (2005)** | US | 4-8 years, TD and SID | A | 31 | - | A | 47 | + |
| KIN | **Ayres (2005)** | US | 4-8 years, TD and SID | A | 46 | - | A | 60 | + |
| FI | **Ayres (2005)** | US | 4-8 years, TD and SID | A | 46 | + | A | 62 | + |
| GRA | **Ayres (2005)** | US | 4-8 years, TD and SID | A | 42 | + | A | 54 | + |
| LTS | **Ayres (2005)** | US | 4-8 years, TD and SID | A | 47 | - | A | 59 | + |
| PRN | **Ayres (2005)** | US | 4-8 years, TD and SID | A | 39 | - | A | 56 | + |
| SWB | **Ayres (2005)** | US | 4-8 years, TD and SID | A | 48 | + | A | 60 | + |

| SIPT Test | Study | Country | Sample | Test-retest Reliability | | | Inter-rater Reliability | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | mq | n | r | mq | n | r |
| MAc | **Ayres (2005)** | US | 4-8 years, TD and SID | A | 45 | + | A | 62 | + |
| DC | **Ayres (2005)** | US | 4-8 years, TD and SID | A | 36 | - | A | 58 | + |
| PPr | **Ayres (2005)** | US | 4-8 years, TD and SID | A | 49 | + | A | 62 | + |
| PrVC | **Ayres (2005)** | US | 4-8 years, TD and SID | A | 48 | + | A | 62 | + |
| CPr | **Ayres (2005)** | US | 4-8 years, TD and SID | A | 51 | - | A | 63 | + |
| SPr | **Ayres (2005)** | US | 4-8 years, TD and SID | A | 47 | + | A | 51 | + |
| OPr | **Ayres (2005)** | US | 4-8 years, TD and SID | A | 49 | + | A | 63 | + |
| BMC | **Ayres (2005)** | US | 4-8 years, TD and SID | A | 45 | + | A | 48 | + |

## Table 2

*Studies Evaluating the Hypothesis-Testing Validity of SIPT*

| SIPT Test | Study | Country | Sample | Convergent Validity | | | Known-groups Validity | | | Developmental Validity | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | mq | n | r | mq | n | r | mq | n | r |
| SV | **Ayres (2005)** | US | 4-8 years, TD, LD, and SID | A | K-ABC = 91 | + | A | see note | + | A | 1997 | + |
| FG | **Ayres (2005)** | US | 4-8 years, TD, LD, and SID | A | K-ABC = 91 | + | A | see note | + | A | 1997 | + |
| MFP | **Ayres (2005)** | US | 4-8 years, TD, LD, and SID | A | K-ABC = 91 | + | A | see note | + | A | 1997 | + |
| KIN | **Ayres (2005)** | US | 4-8 years, TD, LD, and SID | A | K-ABC = 91 | + | A | see note | + | A | 1997 | + |
| FI | **Ayres (2005)** | US | 4-8 years, TD, LD, and SID | A | K-ABC = 91 | + | A | see note | + | A | 1997 | + |
| GRA | **Ayres (2005)** | US | 4-8 years, TD, LD, and SID | A | K-ABC = 91 | + | A | see note | + | A | 1997 | + |
| LTS | **Ayres (2005)** | US | 4-8 years, TD, LD, and SID | A | K-ABC = 91 | + | A | see note | + | A | 1997 | + |
| PRN | **Ayres (2005)** | US | 4-8 years, TD, LD, and SID | A | K-ABC = 91 | + | A | see note | - | A | 1997 | + |
| SWB | **Ayres (2005)** | US | 4-8 years, TD, LD, and SID | A | K-ABC = 91 | + | A | see note | + | A | 1997 | + |
| MAc | **Ayres (2005)** | US | 4-8 years, TD, LD, and SID | A | K-ABC = 91 | + | A | see note | + | A | 1997 | + |
| DC | **Ayres (2005)** | US | 4-8 years, TD, LD, and SID | A | K-ABC = 91 | + | A | see note | + | A | 1997 | + |
| | Cermak & Murray (1991) | US | 5-8 years, TD and LD | A | VMI-Revised, WISC-R Block Design, Primary Visual Motor Test, Rey Osterrieth Complex Figure Test = 39 | + for LD - for TD | A | LD = 21 TD = 18 | + | | | |
| PPr | **Ayres (2005)** | US | 4-8 years, TD, LD, and SID | A | K-ABC = 91 | + | A | see note | + | A | 1997 | + |
| PrVC | **Ayres (2005)** | US | 4-8 years, TD, LD, and SID | A | K-ABC = 91 | + | A | see note | + | A | 1997 | + |
| CPr | **Ayres (2005)** | US | 4-8 years, TD, LD, and SID | A | K-ABC = 91 | + | A | see note | + | A | 1997 | + |
| | Cermak & Murray (1991) | US | 5-8 years, TD and LD | A | VMI-Revised, WISC-R Block Design, Primary Visual Motor Test, Rey Osterrieth Complex Figure Test = 39 | + for LD - for TD | A | LD = 21 TD = 18 | + | | | |
| SPr | **Ayres (2005)** | US | 4-8 years, TD, LD, and SID | A | K-ABC = 91 | + | A | see note | + | A | 1997 | + |
| OPr | **Ayres (2005)** | US | 4-8 years, TD, LD, and SID | A | K-ABC = 91 | + | A | see note | + | A | 1997 | + |
| BMC | **Ayres (2005)** | US | 4-8 years, TD, LD, and SID | A | K-ABC = 91 | + | A | see note | + | A | 1997 | + |

Note. ASD = 7, LD = 195, ABI = 9, ID = 28, SI = 36, Spina Bifida = 21, Reading Disorder = 60, Learning Disorder = 28, CP = 10, TD = 136

## Table 3

*Studies Evaluating the Structural Validity of SIPT*

| SIPT Test | Study | Country | Sample | Structural Validity | | |
|---|---|---|---|---|---|---|
| | | | | mq | n | R |
| Overall | **Ayres (2005)** | US | 4-8 years | A | 1750 | ? - Four factor model (Visuopraxis, Somatopraxis, Vestibular & Somatosensory, Kinesthesia/Motor Accuracy) |
| | **Ayres (2005)** | US | 4-8 years, LD, SID | A | 125 | ? - Five factor model (Bilateral Integration & Sequencing, Praxis on Verbal Command, Somatosensory Processing & Oral Praxis, Visuopraxis, Somatopraxis) |
| | **Ayres (2005)** | US | 4-8 years, TD and LD, SID | A | 293 | ? - Four factor model (Somatopraxis, Visuopraxis, Vestibular Functioning, Somatosensory Processing) |
| | **Mulligan (1998)** | US/Canada | 4-8 years, referred to OT | VG | 10,475 | ? - Four factor model (Visual Perceptual, Bilateral Integration & Sequencing, Dyspraxia, Somatosensory Deficit) with second-order factor (generalized praxis dysfunction) |
| | **Mailloux et al. (2011)** | US | 4-8 years, referred to OT | A | 273 | ? - Four factor model (Visuo-somatodyspraxia, Vestibular and proprioceptive bilateral integration and sequencing, Tactile and visual discrimination, Tactile defensiveness and attention) NOTE: Five item tactile defensiveness scale and SIDngle item attention index included in this analysis; first three factors represent SIPT |
| | **Lai et al. (1996)** | US | 4-20 years (mainly 4-10 years), SID, LD, ADHD, DD | VG | 210 | ? - These five tests represent unidimensional constructs; they can be combined to measure a unitary praxis function |

## Table 4

*Studies Evaluating the Responsiveness of SIPT*

| SIPT Test | Study | Country | Sample | Responsiveness | | |
|---|---|---|---|---|---|---|
| | | | | mq | n | r |
| SV | Kimball (1990) | US | 6 - 8, LD and suspected SID | I | 19 | - |
| FG | Kimball (1990) | US | 7 - 8, LD and suspected SID | I | 19 | - |
| MFP | Kimball (1990) | US | 8 - 8, LD and suspected SID | I | 19 | - |
| KIN | Kimball (1990) | US | 9 - 8, LD and suspected SID | I | 19 | - |
| FI | Kimball (1990) | US | 10 - 8, LD and suspected SID | I | 19 | - |
| GRA | Kimball (1990) | US | 11 - 8, LD and suspected SID | I | 19 | - |
| LTS | Kimball (1990) | US | 12 - 8, LD and suspected SID | I | 19 | + |
| PRN | Kimball (1990) | US | 13 - 8, LD and suspected SID | I | 19 | - |
| SWB | Kimball (1990) | US | 14 - 8, LD and suspected SID | I | 19 | + |
| MAc | Kimball (1990) | US | 15 - 8, LD and suspected SID | I | 19 | - |
| DC | Kimball (1990) | US | 16 - 8, LD and suspected SID | I | 19 | - |
| PPr | Kimball (1990) | US | 17 - 8, LD and suspected SID | I | 19 | + |
| PrVC | Kimball (1990) | US | 18 - 8, LD and suspected SID | I | 19 | - |
| CPr | Kimball (1990) | US | 19 - 8, LD and suspected SID | I | 19 | - |
| SPr | Kimball (1990) | US | 20 - 8, LD and suspected SID | I | 19 | + |
| OPr | Kimball (1990) | US | 21 - 8, LD and suspected SID | I | 19 | - |
| BMC | Kimball (1990) | US | 22 - 8, LD and suspected SID | I | 19 | - |

## Table 5

*Studies Evaluating the Reliability of General Motor Instruments*

| Assessment | Study | Country | Sample | Internal Consistency | | | Test-retest Reliability | | | Inter-rater Reliability | | | Intra-rater Reliability | | | Measurement error | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | mq | n | r | mq | n | r | mq | n | r | mq | n | r | mq | n | r |
| BOT-2 | **Bruininks & Bruininks (2005)*** | USA | 4-21 years | VG | 1520 | + | D | 134 | + | D | 47 | + | | | | A | 1520 | ? |
| | Wuang, Su, and Huang (2012) | | 3-6 years, ID | D | 163 | + | A | 155 | + | | | | | | | A | 141 | - |
| | Wuang & Su (2009) | | 4-12 years, ID | VG | 100 | + | A | 100 | + | | | | | | | A | 100 | + |
| MABC-2 | **Henderson et al. (2012)*** | UK | 3-16 years | | | | D | 60 | + | | | | | | | | | |
| | Valentini et al. (2014) | Brazil | 3-13 years | VG | 844 | + | A | 168 | + | I | 844 | + | D | 844 | + | | | |
| | Wuang, Su, & Su (2012) | Taiwan | 6-12 years, DCD | VG | 144 | + | A | 144 | + | | | | | | | A | 139 | + |
| MABC-2 (Age Band 1 only) | Ellinoudis et al. (2009) | Greece | 3-6 years | VG | 183 | - (manual dexterity)<br>- (aiming & catching)<br>+ (balance) | A | 60 | + | | | | | | | | | |
| | Wuang, Su, & Huang (2012) | Taiwan | 3-6 years, ID | D | 163 | + | A | 155 | + | | | | | | | A | 141 | + |
| | Hua et al. (2013) | China | 3-6 years | D | 1823 | - | A | 184 | + | VG | 184 | + | | | | | | |
| | Serbetar et al. (2019) | Croatia | 3-6 years | | | | A | 174 | + | VG | 36 | + | | | | | | |
| | Hirata et al. (2018) | Japan | 3-6 years | D | 252 | - | | | | | | | | | | | | |
| | Smits-Engelsman et al. (2011) | Netherlands | 3 years | D | 50 | + | A | 28 | + | | | | | | | A | 28 | ? |
| MABC-2 (Age Band 2 only) | Holm et al. (2013) | Norway | 7-9 years | | | | | | | I | 30 | - | I | 29 | - | | | |
| MABC-2 Checklist | Schoemaker et al. (2012) | Dutch | 5-8 years | D | 30 | + | | | | | | | | | | | | |
| | Capistrano et al. (2015) | Brazil | 7-10 years | | | | | | | D | 40 | - | | | | | | |
| PDMS-2 | **Folio & Fewell (2000)** | USA | 0–71 months, US | VG | 2003 | + | | | | | | | | | | A | 2003 | ? |
| | Wuang, Su, & Huang | Taiwan | 3-6 years, ID | D | 163 | + | A | 155 | + | | | | | | | A | 141 | + |
| TGMD-3 | **Ulrich (2019)** | US | 3-11 years, TD and clinical groups | VG | 862 | + | D | 105 | + | | | | | | | A | 862 | ? |
| | Simons et al. (2016) | Belgium | 7-10, ID | VG | 19 | + (locomotor)<br>- (ball skills) | | | | A | 19 | + | A | 19 | + | | | |
| | Webster & Ulrich (2017) | USA | 3-10 years | VG | 807 | + | A | 30 | + | | | | | | | | | |

| Assessment | Study | Country | Sample | Internal Consistency | | | Test-retest Reliability | | | Inter-rater Reliability | | | Intra-rater Reliability | | | Measurement error | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | mq | n | r | mq | n | r | mq | n | r | mq | n | r | mq | n | r |
| | Magistro et al. (2018) | Italy | 3-11, TD and known mental/behavioral disorder | VG | 1075 | + | | | | | | | | | | | | |
| | Maeng et al. (2017) | USA | 3-10 years | | | | | | | VG | 10 | + | I | 10 | + | | | |
| | Valentini et al. (2017) | Brazil | 3-10 years | VG | 597 | + (locomotor) - (ball skills) | D | 128 | + | A | 50 | + | A | 100 | + | | | |
| | Mohammadi et al. (2019) | Iran | 3-10 years | VG | 1600 | + | D | 160 | + | A | 160 | + | A | 160 | + | | | |
| | Estevan et al. (2017) | Spain | 3-10 years | D | 178 | + | | | | | | | | | | | | |
| | Rintala et al. (2017) | Finland | 4-9 years | | | | | | | I | 20 | - | I | 40 | + | | | |
| | Wagner et al. (2017) | Germany | 3-10 years | VG | 189 | + | A | 104 | + | A | 30 | + | A | 30 | + | | | |
| M-FUN | **Miller (2006)** | US | 2.5-7 years, TD and clinical subgroups | VG | 601 | + | D | 28 | + | A | 29 | + | | | | I | 601 | ? |
| COMPS-2 | **Wilson et al. (2000)** | USA | 5-15 years, TD and DCD | VG | 380 | + | A | 48 | + | A | 72 | + | | | | | | |
| SICO | **May-Benson & Teasdale (2021)** | USA | 4-12 years, TD, SPD, and ASD | | | | A | 16 | + | VG | 20 | + | | | | | | |
| SOSI-M | **Blance, Reinoso & Kiefer (2021)** | USA | 5-14, TD, DCD, ADHD, LD, SPD | VG | 1000 | + | D | 31 | + | A | 22 | + | | | | A | 1000 | ? |
| QNST-3R | **Mutti et al. (2017)** | USA | 5-80+ years, TD, ASD, LD, ADHD, unspecified disabilities | VG | 1158 | + for ages 5-11; - for ages 12 - 59; + for ages 60+Z49 | D | 56 (ages 6 - 19 only) | + | | | | | | | | | |
| PBS | **Franjoine et al. (2003)** | USA | 5-15 years, TD and known balance disorders (KBD) | | | | A | 60 | + | VG | 10 | + | | | | | | |
| | Darr et al. (2015) | USA | 2-13 years, TD and known balance disorders | VG | 823 | + | | | | | | | | | | A | 1737 | ? |

## Table 6

*Studies Evaluating Construct Validity of General Motor Instruments based on Hypothesis Testing*

| Assessment | Study | Country | Sample | Convergent Validity | | | Known-groups Validity | | | Predictive Validity | | | Developmental Validity | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | mq | n | r | mq | n | r | mq | n | r | mq | n | r |
| BOT-2 | **Bruininks & Bruininks (2005)** | USA | 4-21 years | A | PDMS-2 = 38<br>TVMS-R = 56 | + | VG | DCD = 50<br>ID = 66<br>Mild ASD = 45<br>TD = Normative Sample | + | | | | VG | 1520 | + |
| | Wuang, Su, and Huang (2012) | Taiwan | 3-6 years, ID | | | | | | | A | PTPS = 141<br>Time = 6 months | + | | | |
| | Lane & Brown (2015) | Australia | 7-16 years | VG | MABC-2 = 50 | + for AB3<br>- for AB2 | | | | | | | | | |
| MABC-2 | **Henderson et al. (2012)** | UK | 3-16 years | | | | I | ASD = 25<br>TD = none | ? | | | | | | |
| | Lane & Brown (2015) | Australia | 7-16 years | VG | BOT-2 = 50 | + for AB3<br>- for AB2 | | | | | | | | | |
| | Griffiths et al. (2017) | Australia | 4-8 years, born < 30 weeks gestation | | | | | | | A | MABC-2<br>96<br>Time = 4 years | + | | | |
| | Valentini et al. (2014) | Brazil | 3-13 years | | | | I | Motor Impairment = 79<br>At-risk = 151<br>TD = 614 | + | | | | | | |
| | Wuang, Su, & Huang (2012) | Taiwan | 3-6 years, ID | | | | | | | A | PTPS = 141<br>Time = 6 months | + | | | |
| | Hua et al. (2013) | China | 3-6 years | VG | PDMS-2 = 184 | + | | | | | | | | | |
| | Capistrano et al. (2015) | Brazil | 7 - 10 years | A | DCDQ-BR = 40 | - | | | | | | | | | |
| | Ulrich (2019) | US | 3-11 years, TD and clinical groups | A | TGMD-3 = 177 | - | | | | | | | | | |
| MABC-2 Checklist | **Henderson et al. (2012)** | UK | 3-16 years | | | | I | ASD = 24 | ? | | | | | | |
| | Schoemaker et al. (2012) | Dutch | 5 - 8 years | A | MABC-2 = 383<br>DCDQ = 130 | + | | | | | | | | | |
| | Capistrano et al. (2015) | Brazil | 7 - 10 years | A | DCDQ-BR = 40 | - | | | | | | | | | |

| Assessment | Study | Country | Sample | Convergent Validity | | | Known-groups Validity | | | Predictive Validity | | | Developmental Validity | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | mq | n | r | mq | n | r | mq | n | r | mq | n | r |
| PDMS-2 | **Folio & Fewell (2000)** | USA | 0 – 71 months, US | A | MSEL:A = 29 | + | D | Physical Disability = 64<br>TD = 2003 | + | | | | VG | 2003 | + |
| | Wuang, Su, & Huang | Taiwan | 3-6 years, ID | | | | | | | A | PTPS = 141<br>Time = 6 months | + | | | |
| | Bruininks & Bruininks (2005)* | USA | 4-21 years | A | PDMS-2 = 38 | + | | | | | | | | | |
| | Van Waelvelde et al. 2007 | Belgium | 4-5 years | A | MABC = 31 | + | | | | | | | | | |
| | Hua et al. (2013) | China | 3-6 years | VG | MABC-2 = 184 | + | | | | | | | | | |
| | Holloway, Long, & Biasini (2019) | US | 4-5 years, ASD | VG | M-FUN = 22 | + | | | | | | | | | |
| TGMD-3 | **Ulrich (2019)** | US | 3-11 years, TD and clinical groups | A | MABC-2 = 177 | - | VG | ASD = 33<br>TD = 34<br>ID = 34 | + | | | | VG | 862 | + |
| | Simons et al. (2016) | Belgium | 7-10, ID | | | | | | | | | | A | 19 | + Locomotor - Ball Skills + Total test |
| | Temple & Foley (2017) | Canada | 8-9 years; 8/277 children with disabilities | | | | | | | | | | D | 277 | + |
| | Wagner et al. (2017) | Germany | 3-10 years | | | | | | | A | German Youth Games, Ball-throwing distance = 91<br>German Youth Games, Sprinting ability = 91<br>Time = 12 months | +<br>- | | | |

| Assessment | Study | Country | Sample | Convergent Validity | | | Known-groups Validity | | | Predictive Validity | | | Developmental Validity | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | mq | n | r | mq | n | r | mq | n | r | mq | n | r |
| M-FUN | **Miller (2006)** | US | 2.5 - 7 years, TD and clinical subgroups | VG | MAP = 15 | + | VG | Visual Motor Delay = 61<br>Fine Motor Delay = 66<br>Gross Motor Delay = 60 | + | | | | | | |
| | Diemand & Case-Smith (2013) | US | 4.5 - 6.5, TD and children receiving OT | VG | DTVP-2 = 40 | + | | | | | | | | | |
| | Holloway, Long, & Biasini (2019) | US | 4-5 years, ASD | VG | PDMS-2 = 22 | + | | | | | | | | | |
| | Rihtman & Parush (2014) | Israel | 2.5 - 7 years | VG | Beery VMI-5 = 30 | + | | | | | | | VG | 267 | + |
| COMPS-2 | **Wilson et al. (2000)** | USA | 5-15 years, TD and DCD | A | BOTMP = 252<br>DCDQ = 202 | + | D | DCD + TD, group n not provided = 315 | + | | | | | | |
| SOSI-M | **Blance, Reinoso & Kiefer (2021)** | USA | 5-14 years, TD, ASD, DCD, LD, ADHD, SPD | VG | SPM = 21<br>SP-2 = 20<br>SIPT PRN = 16 | + | D | ASD = 24<br>DCD = 13<br>LD = 1<br>ADHD = 20<br>SPD = 47<br>Any Disability = 76[1] | + | | | | VG | 1000 | + |
| QNST-3R | **Mutti et al. (2017)** | USA | 5 - 80+ years, TD, ASD, LD, ADHD, unspecified disabilities | VG | CAS = 35<br>Bender-Gestalt II = 26<br>VMI-6 = 27<br>BOT-2 (fine motor) = 19 | + | VG | ADHD = 51<br>LD = 47<br>ASD = 32<br>Dementia = 7<br>Typical = Matched samples | + | | | | D | 1158 | + |
| PBS | **Franjoine et al. (2010)** | USA | 5-15, known balance disorders | | | | | | | | | | A | 641 | + |

[1]Some children in these groups have multiple diagnoses

## Table 7

*Studies Evaluating Structural, Cross-Cultural, and Criterion Validity of General Motor Instruments*

| Assessment | Study | Country | Sample | Structural Validity | | | Cross-cultural Validity/Measurement Invariance | | | Criterion Validity | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | mq | n | r | mq (comparison) | n | r | mq | n | r |
| BOT-2 | **Bruininks & Bruininks (2005)*** | USA | 4-21 years | VG | 1520 | + | | | | | | |
| | Wuang, Lin & Su (2009) | Taiwan | 4 - 18 years, ID | VG | 446 | - | | | | | | |
| | Brown (2019) | Australia | 8-12 years | VG | 123 | - | | | | | | |
| MABC-2 | **Schulz et al. (2011)** | UK | 3-16 years | A | 1172 | + | | | | | | |
| | Valentini et al. (2014) | Brazil | 3-13 years | | | | | | | I | 844 | ? |
| | Psotta & Abdollahipour | Czech Republic | 7-16 years | VG | 1158 | - | | | | | | |
| | Fleurkens-Peeters et al. (2018) | Suriname | 5 years | | | | I (Surinamese vs. UK) | 105 | ? | | | |
| | Ellinoudis et al. (2009) | Greece | 3-6 years | VG | 183 | + | | | | | | |
| | Hua et al. (2013) | China | 3-6 years | A | 1823 | - | | | | | | |
| | Serbetar et al. (2019) | Croatia | 3-6 years | VG | 139 | - | | | | | | |
| | Hirata et al. (2018) | Japan | 3-6 years | VG | 252 | - | I (Japan vs. UK) | 252 | ? | | | |
| | Kita et al. (2016) | Japan | 7-10 years | VG | 132 | + | I (Japan vs. UK) | 132 | ? | | | |
| MABC-2 Checklist | Schoemaker et al. (2012) | Dutch | 5 - 8 years | | | | | | | VG | 383 | - |
| PDMS-2 | **Folio & Fewell (2000)** | USA | 0 – 71 months, US | A | 2003 | + | VG (sex, ethnicity) | 2003 | + | | | |
| | Saraiva et al., 2013 | Portugal | 3-5 years | | | | I (Portugal vs. US) | 540 | ? | | | |
| TGMD-3 | **Ulrich (2019)** | USA | 3-11 years, TD and clinical groups | VG | 862 | + | VG (Sex, race, ethnicity) | 862 | + | | | |
| | Webster & Ulrich (2017) | USA | 3-10 years | VG | 403 | - | | | | | | |
| | Magistro et al. (2018) | Italy | 3-11, TD and known mental/behavioral disorder | | | | VG (TD vs. mental/behavioral disorder) | 1075 | + | | | |
| | Valentini et al. (2017) | Brazil | 3-10 years | VG | 598 | - | VG (sex, age groups) | 598 | + | | | |
| | Estevan et al. (2017) | Spain | 3-10 years | VG | 178 | - | | | | | | |
| | Wagner et al. (2017) | Germany | 3-10 years | VG | 189 | + | VG (sex) | 189 | - | | | |
| M-FUN | **Miller (2006)** | USA | 2.5 - 7 years, TD and clinical subgroups | | | | | | | VG | Visual Motor Delay = 61<br>Fine Motor Delay = 66<br>Gross Motor Delay = 60 | + |

| Assessment | Study | Country | Sample | Structural Validity | | | Cross-cultural Validity/Measurement Invariance | | | Criterion Validity | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | mq | n | r | mq (comparison) | n | r | mq | n | r |
| COMPS | **Wilson et al. (2000)** | USA | 5-15 years, TD and DCD | | | | | | | I | 261 | + |
| SOSI-M | **Blanche, Reinoso & Kiefer (2021)** | USA | 5-14 years, TD, ASD, DCD, LD, ADHD, SPD | | | | VG (sex, ethnicity, hispanic origin, region in US, urban-suburban/rural) | 1000 | + | | | |
| QNST-3R | **Mutti et al. (2017)** | USA | 5 - 80+ years, TD, ASD, LD, ADHD, unspecified disabilities | | | | VG (Sex, Race, Urban-Suburban/Rural) | 1158 | + | A | No disability = 1009<br>Disability (unspecified; ASD or LD) = 64 | + |
| PBS | Darr et al. (2015) | USA | 2-13, TD and known balance disorders | VG | 823 | + | | | | | | |

## Table 8

*Studies Evaluating Reliability of Visual Motor Instruments*

| Assessment | Study | Country | Sample | Internal Consistency | | | Test-retest Reliability | | | Inter-rater Reliability | | | Measurement error | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | mq | n | r | mq | n | r | mq | n | r | mq | n | r |
| Beery VMI | **Beery & Beery (2010)** | US | 1 - 18 years | VG | 750 | + | A | 142 | + | A | 100 | + | I | 1035 | ? |
| | Harvey et al. (2017) | US (Native American Reservation) | 8 - 16 years | | | | VG | 163 | - | VG | 163 | + | | | |
| | Brown (2016) | Australia | 6-8 years, TD | VG | 39 | + | | | | | | | | | |
| DTVP-3 | **Hammill, Pearson & Voress (2014)** | | 4-12 years | VG | 1035 | + | A | 63 | + | A | 30 | + | I | 2160 | ? |
| | Brown (2016) | Australia | 6-8 years, TD | VG | 39 | + | | | | | | | | | |
| TVMS-3 | **Martin (2010)** | US | 3-80+ | VG | 2610 | + | D | 120 | + | A | 64 | + | | | |
| NEPSY-II (VP and DCP) | **Korkman et al. (2007)** | US | 3-16 years, TD and clinical samples (ADHD, LD, ID, ASD, Deaf/HOH, Emotionally Disturbed) | VG | 1450 | + | VG | 160 | DCP (- for multiple age groups) VP (+) | | | | VG | 1450 | ? |

## Table 9

*Studies Evaluating Construct Validity of Visual Motor Instruments based on Hypothesis Testing*

| Assessment | Study | Country | Sample | Convergent Validity | | | Known-groups Validity | | | Developmental Validity | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | mq | n | r | mq | n | r | mq | n | r |
| Beery VMI | **Beery & Beery (2010)** | US | 1 - 18 years | | | | | | | VG | 2758 | + |
| | Brown (2016) | Australia | 6-8 years, TD | VG | DTVP-3 = 29 | - | | | | | | |
| | Pfeiffer et al. (2015) | USA | 5-8 years | A | Minnesota Handwriting Asessment = 207<br>Test of Handwriting Skills Revised = 207 | + | | | | | | |
| | Rihtman & Parush (2014) | Israel | 2.5 - 7 years | VG | M-FUN = 30 | + | | | | | | |
| DTVP-3 | **Hammill, Pearson & Voress (2014)** | USA | 4-12 years | A | Beery VMI-5 = 123<br>TVPS-3 = 123<br>TOSWRF-2 = 31<br>TOSCRF-2 = 25<br>TWS-5 = 48<br>MFaCTs = 61 | + | D | Gifted = 45<br>Heard of Hearing = 12<br>Speech Articulation = 15<br>ADHD = 19<br>Aspberger's disorder = 11<br>LD = 34<br>Physical Impairment = 19<br>Autism = 15<br>Language delay = 23<br>TD = 1035 | + | VG | 1035 | + |
| | Brown (2016) | Australia | 6-8 years, TD | VG | Beery VMI-6 = 29 | - | | | | | | |
| TVMS-3 | **Martin (2010)** | US | 3-80+ | VG | Beery VMI-4 = 26 | + | VG | LD = 66<br>ADHD = 15<br>LD+ADHD = 15<br>Matched TD Groups | + | VG | 2160 | + |
| NEPSY-II (VP and DCP)) | **Korkman et al. (2007)** | US | 3-16 years, TD and clinical samples (ADHD, LD, ID, ASD, Deaf/HOH, Emotionally Disturbed) | A | WISC-IV = 51<br>DAS-II = 242<br>WNV = 62<br>WIAT-II = 81<br>Children's Memory Scale = 43<br>D-KEFS = 49<br>BBCS-3 = 60<br>DSMD = 51<br>ABAS-II = 120<br>Brown ADD Scales for Children and Adolescents = 81<br>CCC-2 = 48 | + | A | ADHD = 55<br>Reading Disorder = 36<br>Mathematics Disorder = 20<br>Language Disorder = 29<br>Intellectual Disability = 20<br>Autistic Disorder = 23<br>Asperger's Disorder = 19<br>Deaf/HOH = 18<br>Emotional Disturbance = 30<br>TD = 1450 | + | | | |

## Table 10

*Studies Evaluating Structural, Cross-Cultural, Criterion Validity, and Responsiveness of Visual Motor Instruments*

| Assessment | Study | Country | Sample | Structural Validity | | | Cross-cultural Validity/Measurement Invariance | | | Criterion Validity | | | Responsiveness | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | mq | n | r | mq (comparison) | n | r | mq | n | r | mq | n | r |
| VMI-6 | **Beery & Beery (2010)** | US | 1 - 18 years | I | 2758 | + | VG (Sex) | 2758 | + | | | | | | |
| | Brown (2016) | Australia | 6-8 years | | | | | | | | | | | | |
| | Pfeiffer et al. (2015) | USA | 5-8 years | | | | | | | | | | D | 207 | ? |
| DTVP-3 | **Hammill, Pearson & Voress (2014)** | USA | 4-12 years | VG | 1035 | + | VG (Sex and Race) | 1035 | + | A | Visual Perception Impairment = 10 No Impairment = 65 | + | | | |
| | Brown (2016) | Australia | 6-8 years | | | | | | | | | | | | |
| TVMS-3 | **Martin (2010)** | USA | 3-80+ | VG | 2160 | + | | | | | | | | | |

## Table 11

*Studies Evaluating Reliability of Praxis Instruments*

| Assessment | Study | Country | Sample | Internal Consistency | | | Test-retest Reliability | | | Inter-rater Reliability | | | Intra-rater Reliability | | | Measurement error | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | mq | n | r | mq | n | r | mq | n | r | mq | n | r | mq | n | r |
| TIP | Lane et al. (2014) | US | 3-5 years | | | | A | 16 | + | VG | 19 | + | | | | | | |
| PIPS | Vanvuchelen, Roeyers & de Weerdt (2011a) | Belgium | 12 - 59 months, TD and ASD | VG | 298 | + | | | | | | | | | | | | |
| | Vanvuchelen, Roeyers & de Weerdt (2011b) | Belgium | 18 - 59 months | | | | A | 56 | + | VG | 42 | + | VG | 21 | + | VG | 119 | ? |
| | Vanvuchelen & Vochten, 2010 | Belgium | 13 - 58 months, ID | | | | | | | VG | 44 | + | | | | VG | 44 | ? |
| NEPSY-II (FT, IH, MM) | **Korkman et al. (2007)** | US | 3-16 years, TD and clinical samples (ADHD, LD, ID, ASD, Deaf/HOH, Emotionally Disturbed) | VG | 1450 | + | A (FT only) | 165 | + | | | | | | | VG | 1450 | ? |

## Table 12

*Studies Evaluating Construct Validity of Praxis Instruments based on Hypothesis Testing*

| Assessment | Study | Country | Sample | Convergent Validity | | | Known-groups Validity | | | Developmental Validity | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | mq | n | r | mq | n | r | mq | n | r |
| NEPSY-II (FT, IH, MM) | **Korkman et al. (2007)** | US | 3-16 years, TD and clinical samples (ADHD, LD, ID, ASD, Deaf/HOH, Emotionally Disturbed) | VG (FT only) | WISC-IV = 51<br>DAS-II = 242<br>WNV = 62<br>WIAT-II = 81<br>Children's Memory Scale = 45<br>D-KEFS = 49<br>BBCS-3 = 60<br>DSMD = 51<br>ABAS-II = 120<br>Brown ADD Scales for Children and Adolescents = 81<br>CCC-2 = 48 | + | VG (FT only) | ADHD = 55<br>Reading Disorder = 36<br>Mathematics Disorder = 20<br>Language Disorder = 29<br>Intellectual Disability = 20<br>Autistic Disorder = 23<br>Asperger's Disorder = 19<br>Deaf/HOH = 18<br>Emotional Disturbance = 30<br>TD = 1450 | + | | | |
| PIPS | Vanvuchelen, Roeyers & de Weerdt (2011a) | Belgium | 12 - 59 months, TD and ASD | | | | | | | A | 545 | + |

## Table 13

*Studies Evaluating Structural, Cross-Cultural, Criterion Validity, and Responsiveness of Visual Motor Instruments*

| Assessment | Study | Country | Sample | Structural Validity | | | Cross-cultural Validity/Measurement Invariance | | | Criterion Validity | | | Responsiveness | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | mq | n | r | mq (comparison) | n | r | mq | n | r | mq | n | r |
| PIPS | Vanvuchelen, Roeyers & de Weerdt (2011b) | Belgium | 12 - 59 months, TD and ASD | A | 498 | + | | | | | | | | | |

## Table 14

*Studies Evaluating Reliability of Sensory Perception Instruments*

| Assessment | Study | Country | Sample | Internal Consistency | | | Test-retest Reliability | | | Measurement error | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | mq | n | r | mq | n | r | mq | n | r |
| COP | **Blanche et al. (2012)** | US | | | | | | | | | | |
| MVPT-4 | **Colarusso & Hammill (2015)** | US | 4-14 years | VG | 2160 | + | A | 60 | + | I | 2160 | + |
| | Kose et al. (2019) | Turkey | 7-10 years, TD and LD | | | | | | | | | |
| TVPS-4 | **Martin (2017)** | US | 5-21 years | VG | 1790 | + for all subtests except "SEQ" | A | 71 | + | I | 1790 | ? |
| NEPSY-II (AW, GP, PP, RF) | **Korkman et al. (2007)** | US | 3-16 years, TD and clinical samples (ADHD, LD, ID, ASD, Deaf/HOH, Emotionally Disturbed) | VG (AW, GP, and PP only) | 1450 | + | VG (AW, GP, and PP only) | 160 | AW (-) GP (-) PP (+) | VG (AW, GP, and PP only) | 1450 | ? |

## Table 15

*Studies Evaluating Construct Validity of Sensory Perception Instruments based on Hypothesis Testing*

| Assessment | Study | Country | Sample | Convergent Validity | | | Known-groups Validity | | | Developmental Validity | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | mq | n | r | mq | n | r | mq | n | r |
| COP | **Blanche et al. (2012)** | US | | A | 24 | + | I | Known Proprioceptive Problems = 24<br>TD = 20 | + | | | |
| MVPT-4 | **Colarusso & Hammill (2015)** | US | 4-14 years, TD and clinical groups | VG | TVPS -3 = 27 | + | A | DD = 38<br>Acquired Brain Injury = 48<br>LD = 51<br>TD = 2160 | + | VG | 1870 | + |
| | Kose et al. (2019) | Turkey | 7-10 years, TD and LD | | | | A | TD = 48<br>LD = 48 | + | | | |
| | Martin (2017) | US | 5-21 years | VG | TVPS-4 = 32 | + | | | | | | |
| TVPS-4 | **Martin (2017)** | US | 5-21 years | VG | MVPT-4 = 32 | + | A | ADHD = 43<br>ASD = 43<br>TD = 29 | + | VG | 1790 | + |
| NEPSY-II (AW, GP, PP, RF) | **Korkman et al. (2007)** | US | 3-16 years, TD and clinical samples (ADHD, LD, ID, ASD, Deaf/HOH, Emotionally Disturbed) | A (AW, GP, and PP only) | WISC-IV = 51<br>DAS-II = 242<br>WNV = 62<br>WIAT-II = 81<br>Children's Memory Scale = 43<br>D-KEFS = 49<br>BBCS-3 = 60<br>DSMD = 51<br>ABAS-II = 120<br>Brown ADD Scales for Children and Adolescents = 81<br>CCC-2 = 48 | + | A (AW, GP, and PP only) | ADHD = 55<br>Reading Disorder = 36<br>Mathematics Disorder = 20<br>Language Disorder = 29<br>Intellectual Disability = 20<br>Autistic Disorder = 23<br>Asperger's Disorder = 19<br>Deaf/HOH = 18<br>Emotional Disturbance = 30<br>TD = 1450 | + | | | |

## Table 16

*Studies Evaluating Structural, Cross-Cultural, Criterion Validity, and Responsiveness of Sensory Perception Instruments*

| Assessment | Study | Country | Sample | Structural Validity | | | Cross-cultural Validity/Measurement Invariance | | | Criterion Validity | | | Responsiveness | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | mq | n | r | mq (comparison) | n | r | mq | n | r | mq | n | r |
| COP | **Blanche et al. (2012)** | US | | A | 130 | + | | | | | | | | | |
| MVPT-4 | **Colarusso & Hammill (2015)** | US | 4-14 years* | | | | VG (sex, race, ethnicity) | 2160 | + | | | | | | |
| | Kose et al. (2019) | Turkey | 7-10 years, TD and LD | | | | | | | A | 96 | + | | | |
| TVPS-4 | **Martin (2017)** | US | 5-21 years | VG | 1790 | + | VG (sex, urban/rural, race) | 1790 | + | | | | | | |

## Table 17

*Studies Evaluating Reliability of Sensory Reactivity Instruments*

| Assessment | Study | Country | Sample | Internal Consistency | | | Test-retest Reliability | | | Inter-rater Reliability | | | Measurement error | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | mq | n | r | mq | n | r | mq | n | r | mq | n | r |
| SP-2 | **Dunn (2014)** | US | 3-14, TD and clinical groups | VG | 697 | + | D | 113 | + | A | 82 | + | VG | 697 | ? |
| SSP-2 | **Dunn (2014)** | US | 3-14, TD and clinical groups | VG | 697 | + | D | 113 | + | | | | VG | 697 | ? |
| | Chojnicka et al. (2019) | Poland | 3-14 years, TD, ASD, and DD | VG | 1230 | + | A | 24 | + | | | | | | |
| SPM | **Parham et al. (2007)** | US | 5-12 years | VG | 1396 | + | A | 77 | + | | | | VG | 1396 | ? |
| | Brown et al. (2010b) | Australia | 5 - 10 years | VG | 60 | + | | | | A (Home form only) | 60 | - | | | |
| | Lai et al. (2011) | Hong Kong | 5-12 years, TD and ASD | VG | 542 | + | A | SPM-Home = 28 SPM-Classroom = 21 | + | | | | | | |
| | Alkhalifah et al. (2020) | Saudi Arabia | 2-12 years, TD and ASD | VG | 117 | + | | | | | | | | | |

## Table 18

*Studies Evaluating Construct Validity of Sensory Reactivity Instruments based on Hypothesis Testing*

| Assessment | Study | Country | Sample | Convergent Validity | | | Known-groups Validity | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | mq | n | r | mq | n | r |
| SP-2 | **Dunn (2014)** | US | 3-14, TD and clinical groups | A | BASC-2 = 51<br>SSIS = 56<br>VABS-2 = 45 | +<br>+<br>- | A | DD = 11<br>ASD = 78<br>ADHD = 96<br>ADHD/ASD = 24<br>LD = 45<br>G&T = 18<br>All vs. TD = 697 | + |
| SSP-2 | **Dunn (2014)** | US | 3-14, TD and clinical groups | A | BASC-2 = 51<br>SSIS = 56<br>VABS-2 = 45 | +<br>+<br>- | A | DD = 11<br>ASD = 78<br>ADHD = 96<br>ADHD/ASD = 24<br>LD = 45<br>G&T = 18<br>All vs. TD = 697 | + |
| | Chojnicka et al. (2019) | Poland | 3-14 years, TD, ASD, and DD | A | SCQ = 1230 | + | VG | ASD = 310<br>DD = 264<br>TD = 656 | + |
| SPM | **Parham et al. (2007)** | US | 5-12 years | A | SP = 182 | + | VG | Children receiving OT services = 182 | + |
| | Brown et al. (2010a) | Australia | 5 - 10 years | VG | SP = 30<br>SPSC = 19 | + | | | |
| | Lai et al. (2011) | Hong Kong | 5-12 years, TD and ASD | A | Chinese SP = 44 | + | D | TD = 100<br>ASD = 100 | + |
| | Alkhalifah et al. (2020)1 | Saudi Arabia | 2-12 years, TD and ASD | | | | VG | ASD = 93<br>TD = 24 | + |
| | Hansen & Jirikowic (2013) | USA | 5-11, TD and fetal alcohol spectrum disorder (FASD) | A | SSP = 23 | + | A | TD = 12<br>FASD = 11 | + |

**Table 19**

*Studies Evaluating Structural, Cross-Cultural, Criterion Validity, and Responsiveness of Sensory Perception Instruments*

| Assessment | Study | Country | Sample | Structural Validity | | | Cross-cultural Validity/Measurement Invariance | | | Criterion Validity | | | Responsiveness | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | mq | n | rating | mq | n | rating | mq | n | rating | mq | n | rating |
| SSP-2 | Chojnicka et al. (2019) | Poland | 3-14 years, TD, ASD, and DD | A | 1230 | - | | | | | | | | | |
| | Parks et al. (2020) | Canada | 1 - 22 years, ASD and ADHD | VG | 571 | + | VG (ASD vs. ADHD) | 571 | - | | | | | | |
| Sensory Processing Measure (SPM) | **Parham et al. (2007)** | US | 5-12 years | A | 1051 | + | | | | D | 1084 | + | | | |
| | Lai et al. (2011) | Hong Kong | 5-12 years, TD and ASD | A | 542 | - | | | | | | | | | |
| | Alkhalifah et al. (2020) | Saudi Arabia | 2-12 years, TD and ASD | | | | | | | VG (Home form only) | 117 | + | | | |
| SEQ 3.0 | **Ausderau et al. (2014)** | USA | 2-12 years, ASD | VG | 1407 | + | VG (sex, age) | 1407 | + | | | | | | |

# APPENDIX C

## Table 1

*Study Participants by Sex, Race/Ethnicity, and Living Area*

| Country | Sex | | Race/Ethnicity | | | | | | | Living Area | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Male | Female | Native American | Asian | Black Non-Hispanic | White Non-Hispanic | Hispanic | Other/ Unknown | Missing | Urban | Suburban | Rural | Missing | |
| Australia/New Zealand ($n$ = 25) | | | | | | | | | | | | | | |
| Australia | 9 | 6 | 0 | 1 | 0 | 13 | 0 | 1 | 0 | 14 | 1 | 0 | 0 | **15** |
| New Zealand | 4 | 6 | 0 | 0 | 0 | 4 | 0 | 6 | 0 | 8 | 1 | 1 | 0 | **10** |
| Eastern Asia ($n$ = 530) | | | | | | | | | | | | | | |
| China | 124 | 121 | 0 | 240 | 0 | 0 | 0 | 5 | 0 | 148 | 49 | 10 | 38 | **245** |
| Hong Kong | 5 | 14 | 0 | 19 | 0 | 0 | 0 | 0 | 0 | 19 | 0 | 0 | 0 | **19** |
| South Korea | 23 | 20 | 0 | 43 | 0 | 0 | 0 | 0 | 0 | 15 | 0 | 0 | 28 | **43** |
| Macau | 6 | 5 | 0 | 11 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 6 | **11** |
| Malaysia | 21 | 31 | 0 | 52 | 0 | 0 | 0 | 0 | 0 | 36 | 16 | 0 | 0 | **52** |
| Singapore | 16 | 12 | 0 | 28 | 0 | 0 | 0 | 0 | 0 | 22 | 3 | 0 | 3 | **28** |
| Taiwan | 67 | 65 | 0 | 132 | 0 | 0 | 0 | 0 | 0 | 19 | 20 | 3 | 90 | **132** |
| Eastern Europe ($n$ = 196) | | | | | | | | | | | | | | |
| Bulgaria | 5 | 6 | 0 | 0 | 0 | 10 | 0 | 1 | 0 | 10 | 1 | 0 | 0 | **11** |
| Croatia | 1 | 4 | 2 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 3 | 2 | 0 | **5** |
| Czech Republic | 16 | 19 | 0 | 0 | 0 | 35 | 0 | 0 | 0 | 25 | 7 | 3 | 0 | **35** |
| Estonia | 4 | 6 | 0 | 0 | 0 | 10 | 0 | 0 | 0 | 5 | 4 | 0 | 1 | **10** |
| Lithuania | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | **1** |
| Romania | 22 | 39 | 0 | 0 | 0 | 60 | 0 | 1 | 0 | 36 | 16 | 7 | 2 | **61** |

| Country | Sex | | Race/Ethnicity | | | | | | | Living Area | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Male | Female | Native American | Asian | Black Non-Hispanic | White Non-Hispanic | Hispanic | Other/ Unknown | Missing | Urban | Suburban | Rural | Missing | |
| Russia | 22 | 20 | 0 | 0 | 1 | 41 | 0 | 0 | 0 | 30 | 5 | 6 | 1 | **42** |
| Slovenia | 10 | 17 | 0 | 0 | 0 | 26 | 0 | 1 | 0 | 1 | 14 | 12 | 0 | **27** |
| Turkey | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 3 | 0 | 1 | 0 | **4** |
| Latin America (*n* = 395) | | | | | | | | | | | | | | |
| Argentina | 42 | 64 | 4 | 0 | 0 | 2 | 98 | 1 | 1 | 86 | 19 | 0 | 1 | **106** |
| Brazil | 24 | 47 | 3 | 1 | 9 | 49 | 2 | 5 | 2 | 65 | 3 | 3 | 0 | **71** |
| Colombia | 34 | 49 | 0 | 0 | 0 | 1 | 79 | 3 | 0 | 50 | 19 | 13 | 1 | **83** |
| Costa Rica | 0 | 5 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 1 | 4 | 0 | 0 | **5** |
| Dominican Republic | 7 | 4 | 0 | 0 | 0 | 1 | 10 | 0 | 0 | 6 | 5 | 0 | 0 | **11** |
| El Salvador | 11 | 3 | 0 | 0 | 0 | 0 | 14 | 0 | 0 | 8 | 2 | 2 | 2 | **14** |
| Guatemala | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | **1** |
| Haiti | 2 | 1 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | **3** |
| Mexico | 40 | 61 | 0 | 0 | 0 | 1 | 100 | 0 | 0 | 73 | 13 | 12 | 3 | **101** |
| Middle Eastern and North Africa (*n* = 23) | | | | | | | | | | | | | | |
| Israel | 5 | 8 | 0 | 0 | 0 | 10 | 0 | 0 | 3 | 2 | 7 | 0 | 4 | **13** |
| United Arab Emirates | 4 | 6 | 0 | 0 | 0 | 10 | 0 | 0 | 0 | 10 | 0 | 0 | 0 | **10** |
| North America (*n* = 121) | | | | | | | | | | | | | | |
| Canada | 17 | 24 | 1 | 0 | 0 | 39 | 0 | 1 | 0 | 22 | 12 | 6 | 1 | **41** |
| United States | 33 | 47 | 1 | 5 | 2 | 64 | 4 | 3 | 1 | 28 | 32 | 19 | 1 | **80** |
| South Asia (*n* = 181) | | | | | | | | | | | | | | |
| India | 40 | 57 | 0 | 97 | 0 | 0 | 0 | 0 | 0 | 60 | 24 | 13 | 0 | **97** |
| Pakistan | 40 | 44 | 0 | 80 | 0 | 0 | 0 | 0 | 4 | 4 | 22 | 0 | 58 | **84** |
| Sub-Saharan Africa (*n* = 340) | | | | | | | | | | | | | | |
| Namibia | 18 | 16 | 0 | 0 | 6 | 15 | 0 | 13 | 0 | 34 | 0 | 0 | 0 | **34** |

| Country | Sex | | Race/Ethnicity | | | | | | | Living Area | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Male | Female | Native American | Asian | Black Non-Hispanic | White Non-Hispanic | Hispanic | Other/ Unknown | Missing | Urban | Suburban | Rural | Missing | |
| South Africa | 143 | 163 | 0 | 10 | 163 | 38 | 2 | 86 | 7 | 188 | 72 | 29 | 17 | **306** |
| Western Europe ($n$ = 598) | | | | | | | | | | | | | | |
| Austria | 26 | 25 | 0 | 0 | 0 | 19 | 0 | 21 | 11 | 11 | 11 | 14 | 15 | **51** |
| Denmark | 12 | 18 | 0 | 0 | 0 | 25 | 0 | 5 | 0 | 9 | 16 | 3 | 2 | **30** |
| England | 54 | 70 | 0 | 0 | 0 | 14 | 0 | 109 | 1 | 33 | 65 | 23 | 3 | **124** |
| Finland | 5 | 10 | 0 | 0 | 0 | 6 | 0 | 9 | 0 | 0 | 5 | 0 | 10 | **15** |
| Germany | 26 | 40 | 0 | 0 | 0 | 49 | 1 | 16 | 0 | 27 | 22 | 15 | 2 | **66** |
| Iceland | 4 | 6 | 0 | 0 | 0 | 10 | 0 | 0 | 0 | 10 | 0 | 0 | 0 | **10** |
| Ireland | 7 | 1 | 0 | 0 | 0 | 8 | 0 | 0 | 0 | 4 | 4 | 0 | 0 | **8** |
| Netherlands | 31 | 34 | 0 | 2 | 0 | 31 | 0 | 31 | 1 | 16 | 31 | 2 | 16 | **65** |
| Norway | 7 | 6 | 0 | 0 | 0 | 9 | 2 | 2 | 0 | 4 | 8 | 0 | 1 | **13** |
| Poland | 11 | 19 | 0 | 0 | 0 | 30 | 0 | 0 | 0 | 12 | 9 | 9 | 0 | **30** |
| Portugal | 10 | 6 | 0 | 0 | 0 | 16 | 0 | 0 | 0 | 7 | 7 | 2 | 0 | **16** |
| Scotland | 9 | 10 | 0 | 0 | 0 | 0 | 0 | 17 | 2 | 4 | 9 | 6 | 0 | **19** |
| Spain | 47 | 59 | 0 | 0 | 0 | 59 | 41 | 6 | 0 | 49 | 36 | 13 | 8 | **106** |
| Sweden | 4 | 7 | 0 | 0 | 0 | 11 | 0 | 0 | 0 | 2 | 4 | 5 | 0 | **11** |
| Switzerland | 13 | 13 | 0 | 0 | 0 | 11 | 0 | 15 | 0 | 4 | 5 | 16 | 1 | **26** |