Evaluating the Psychometric Properties of the 18-month "Ages and Stages Questionnaires" using a Canadian Sample: Making Inferences about Decision Consistency from Item and Subscale-Level Data

by

Ekaterina Chudnovskaya

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Education
in
Measurement, Evaluation and Cognition

Department of Educational Psychology
University of Alberta

.

Abstract

Developmental screening tools, such as "Ages and Stages Questionnaires" (ASQ) are an important addition to the pediatric care to identify developmental delays at critical age periods. ASQ questionnaires have demonstrated good psychometric properties in the US context; however sample-dependent methods were used to establish this evidence and set cut-off scores. The purpose of this study was to evaluate the psychometric properties of the ASQ for 18 month children in the new context – primary care population in a Western Canadian community. A combination of classical test theory and non-parametric item response theory methods for item and subscale analysis were used to make inferences about potential consistency of classifications with the original cut-off scores. Results indicate that (a) cut-off locations do not match the original distribution for most subscales; (b) high probability of misclassification exists for subscales, despite acceptable internal consistency; (c) item difficulty ranges from low to acceptable, but contributes to low discrimination around cut-offs for some subscales; (d) all domains provide lower precision and discrimination at higher ability levels, thus increasing potential of misclassification for children not clearly at risk. Implications of these findings for research and tool use, including scoring and interpretation of results, are discussed.

Preface

This thesis is an original work by Ekaterina Chudnovskaya. The research project, of which this thesis is a part, received research ethics approval from the University of Alberta Research Ethics Board, Project Name 'Psychometric Properties of ASQ-18 in Canadian context', # MS1_Pro00039631, July 29th, 2013.

Acknowledgements

This project was made possible thanks to the invaluable input of many people: faculty members, fellow students, friends and family. I would like to use this opportunity to express my sincere gratitude to all of them.

To my supervisor, Dr. Mark Gierl, thank you for sparking my interest in psychometric modelling and inspiring me to complete this project with your own research and teaching excellence. I especially appreciate your continuous support at all stages of this project, constructive criticism, patience and encouragement. To my mentors in CRAME, Dr. Cheryl Poth and Dr. Todd Rogers, thank you for your guidance and teaching me about the fields of educational measurement and evaluation through coursework, individual feedback and advice during the formative years of my program. I am especially grateful to Dr. Poth for valuable teaching and evaluation experience that made me a stronger researcher and writer, and for providing feedback on my thesis as committee member. To my mentor at CUP and another committee member, Dr. Rebecca Gokiert, thank you for making this study possible by granting me access to your data, and for sharing your child development expertise at various stages of the thesis completion process. To past and present fellow CRAMErs Oksana Babenko, Amin Mousavi, Qi Guo, Ulemu Luhanga, Man-Wai Chu, Dorothy Pinto and Maria-Clara Bustos Gomez, thank you for your suggestions, help and support.

Last but not least, I am indebted to my family and my closest friends for being there for me in the best and worst moments. This work is dedicated to them.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The first years of life are a critical period for child development during which time the interplay of genetics and environment influence the emergence and development of skills and competencies throughout the lifespan (Pool & Hourcade, 2011). Child development progresses along a continuum of acquiring more and more complex skills and competencies, from reflexive behaviours to integrated functions (Aylward, 2009). Despite the velocity of change and intra-individual variation in skill acquisition, the emergence of developmental milestones progresses in a somewhat predictable manner (Eliot, 1999), with many of the functions fully developed by age six. This combination of known critical periods with expected individual variation opens up fixed "windows of opportunity", in which inconsistencies or discontinuities in development can be identified and addressed through early intervention (Aylward, 1997; Limbos & Joyce, 2011). Should such opportunity be missed, an initially asymptomatic child may proceed to develop a physical and/or cognitive disability, impeding their learning, health and social functioning later in life (Limbos & Joyce, 2011).

Background to the Problem

The term "developmental delay" describes the situation of a child failing to meet an age-appropriate milestone at the latest possible age (Accardo & Whitman, 2005). In North America, 5% to 20% of children are expected to develop at least one type of a delay or disorder before they turn 18 (American Academy of Pediatrics, 2006). Other estimates show that about 25% of the children fail developmental screening, and about 10% of these children are later diagnosed with a disability (Aylward, 1997), meaning that they display

substantial functional limitations in major life activities, such as independent living, ability to learn or economic self-sufficiency (Accardo & Whitman, 2005).

According to the American Academy of Pediatrics (American Academy of Pediatrics, 2006), the combination of routine observations with the use of a brief developmental screening tool at critical periods in child development allows for the most accurate identification of developmental delays in children and increases their access to diagnostic evaluation and early intervention. Such screening instruments can be completed either by trained staff or by parents and are typically short, easy to administer, score and interpret and fit into a busy pediatric practice or a community health centre (Gokiert et.al., 2014; Berger, Hopkins, Bae, Hella, & Strickland, 2010). However, these requirements are known to negatively impact sensitivity of the tools resulting in a large number of over-referrals (Glascoe, 2001; Glascoe, 2005).

"Ages and Stages Questionnaires" (ASQ: Squires, Twombly, Bricker, & Potter, 2009) is a parent-completed screening tool used for developmental screening in the US and other countries, including Canada (Rydz et.al., 2006; Simard, Luu, & Gosselin, 2012) . It is a system of age-specific questionnaires aimed at identifying children at risk for delays between the ages of 1 and 66 months. Each questionnaire in this system elicits children's performance on important milestones across the main developmental domains (e.g. communication, motor skills etc.; see literature review for details). Each age interval also has two specific cut-off points separating children into the at-risk group (immediate referral to assessment or services) and the monitoring zone (need for follow-up and some specific activities).

The ASQ is one of the most studied tools among general developmental screening questionnaires (Macy, 2012) and has established psychometric properties for the US population (Marks & La Rosa, 2012). However, validation studies on the third version of the ASQ have focused mostly on the accuracy of decisions, summarized across multiple age periods. Moreover, reliability and some aspects of validity evidence have only been reviewed for specific age periods (Squires et al., 2009). The findings regarding the use of American norms in different contexts are contradictory (e.g., (Kerstjens et al., 2009; Frisk et al., 2009). Item quality could be a reason for high rates of over-referrals in some contexts, especially because item difficulty has not been verified empirically for most periods. Given that some periods, including the 18-month period, were recognized as critical for identifying delays and providing early interventions (American Academy of Pediatrics, 2006), and given that both methods used for setting cut-off scores and validating ASQ scales were sample specific (Squires et al., 2009), item functioning for those specific scales need to be scrutinized in more detail when used in a population with potentially different characteristics.

Purpose and Overview of the Study

The purpose of the present study is to evaluate how well the ASQ for the 18-month age interval, henceforth referred to as "ASQ-18", can consistently identify children at risk for developmental delays in a primary care practice in a Canadian urban community. This context is associated with lower local prevalence of delays in children, compared to the national US normative sample. Also, Canadian data were collected for the purpose of surveillance rather than for standard setting and cut-off evaluation, as no secondary screen was administered to the children. Under these limitations, the inferences about decision

consistency will be made from the domain-level reliability measures, from the review of item difficulty and discrimination, as well as from the amount of information the domain scores provide at the ability levels around the cut-off scores. While the proposed methods cannot be taken as a substitute for a full decision accuracy study, they still provide useful information about scale functioning in a different population.

This paper is organized as follows: Chapter 2 provides a review of the scholarly literature on developmental screening, in general, and the ASQ-18, in particular. Chapter 3 outlines the methods used in this study. Chapter 4 describes the results for domain subscales and items. Chapter 5 summarizes the findings, explains limitations and provides directions for future research.

Chapter 2: Literature Review

This chapter describes (1) the context of and psychometric evidence of developmental screening tools and (2) content and psychometric properties for the Ages and Stages Questionnaires. The first section outlines the rationale for developmental screening and reviews the evidence on the psychometric properties that should characterise a screening tool including not only diagnostic accuracy and reliability, but also analysis of item quality using non-parametric item response theory. The second section focuses on the content and administration of the ASQ-18, describes the psychometric properties demonstrated on the US normative (or standardization) sample and briefly discusses findings from non-US studies. The chapter concludes with a statement of the problem and the research question.

The Context of Developmental Screening

Developmental screening rationale. During the first five years of life, changes in a child's brain and nervous system as well as psychomotor development should progress in a somewhat predictable order in four large domains (Aylward, 1997; Rydz, Shevell, Majnemer, & Oskoui, 2005; Bellman, Byrne, & Sege, 2013).

1. Motor skills, including gross motor (sitting, walking and changing position) and fine motor (manipulation of objects with fingers);
2. Speech and language (receptive and expressive language skills, articulation of sounds, and symbolic use of gestures);

3. Personal-social and emotional development (prosocial behaviour and self-help skills for daily living); and
4. Cognitive development (problem solving through "intuition, perception and verbal and non-verbal reasoning" (Rydz et.al., 2005, p.5).

Some of the key tasks, called developmental milestones (e.g. walking, imitating adults' actions, first words), typically manifest themselves at a certain age, although individual variation can be large for some milestones (Eliot, 1999; Bellman et.al., 2013). If achievement of the milestones is significantly delayed in one or more domain or if there is a marked discontinuity between skills emerging in various domains, then a delayed course of development is suspected.

In North America, 5% to 20% of children are expected to develop a developmental delay or disorder by the age of 18 (American Academy of Pediatrics, 2006). For infants and toddlers under 2 years of age, the prevalence differs: from 1%-2% for global delays that are expected to affect several domains at once and from 5%-10% for specific disorders affecting one domain, especially language development (Sonnander, 2000; Glascoe, 2005; Bellman et al., 2013). A significant body of research has established that effectiveness of timely identification of delays and referrals to early interventions result in positive outcomes for children's health (Glascoe & Dworkin, 2008; Pool & Hourcade, 2011). At the same time, an early delay that is not identified has the potential to lead to significant learning difficulties later in life due to the cumulative nature of competency development (Aylward, 1997).

The American Academy of Pediatrics recommends using brief developmental screening instruments at key age intervals, given a child's risk and medical history

(American Academy of Pediatrics, 2006), as the combination of screening and regular pediatric surveillance has proven to be more effective in identifying delays, than the clinical judgement alone (Marks & LaRosa, 2012). The age of 18 months is one of the most critical periods for screening because delays in most major areas of development are detectable and effective early intervention approaches exist (American Academy of Pediatrics, 2006). It is important to remember, however, that brief screening measures are not diagnostic tools: they indicate risk of a delay but should be followed by an in-depth assessment in the areas of concern using multiple lines of evidence (Berger at al., 2010). The results of the screening test should also be interpreted by clinicians in the context of parents' concern, child's medical history and earlier observations (Aylward, 2009).

Psychometric properties of developmental screening tools. According to the screening algorithm of the American Academy of Pediatrics (American Academy of Pediatrics, 2006), repeated in the guidelines for practitioners (e.g. Marks & LaRosa, 2012), the psychometric properties of a screening tool required for the accurate interpretation of test scores include up to date norms, derived from the representative sample, reliable scores and valid scores, defined as the "ability to discriminate between a child at a determined level of risk for delays… and the rest of the population" (American Academy of Pediatrics, 2006, pp.416-417). At the same time, the practical constraints of developmental screening require the tools to be short, quick to complete, low cost for administration, minimal training for administration and scoring by the staff and well integrated into the clinical context (Rydz et.al., 2005; Berger at al., 2010). These practical requirements have an impact on the consistency of findings, as brief tools are expected to have high portion of error in the scores because of their length (Emons, Sijtsma, & Meijer, 2007), while the

ongoing need to balance sensitivity and specificity poses a risk of over-or underidentification (Pool & Hourcade, 2011). In addition, developmental ability can be conceptualized as a continuum, with children falling above the at-risk cut-off still underperforming compared to average abilities of their peers (Aylward, 1997; Glascoe, 2001; Bellman et al., 2013). Consequently, screening tools need to be able to not only identify "presence" or "absence" of risk, but also rank children on the continuum for a range of the developmental ability (see Santor, 2005 for a related argument for mental health screening). Psychometric studies on developmental screening tests, as well as the recommendations for clinicians in the American Academy of Pediatrics screening algorithm (American Academy of Pediatrics, 2006) tend to treat developmental ability as a categorical variable referring to the presence or absence of risk.

Diagnostic accuracy and reliability of developmental screeners. A recent review of developmental screening tests conducted by Macy (2012) indicated that the validity evidence based on external criteria - correlations between concurrently administered scales, conditional probabilities, and, to a lesser extent, prediction of later functioning – are the most commonly reported indicator. In comparison, there are half as many studies reporting reliability evidence (Macy, 2012). For reliability, internal consistency is reported less often than test-retest and inter-rater reliability. One possible reason for this reliability outcome is the nature of development. Development should be judged by the ability of a child to perform complex functional tasks at a specific age which requires the integration of skills from several domains (Aylward, 2009). In some recommendations, reliability is not specified (e.g. Drotar, Stancin, Dworkin, Sices, & Wood, 2008; Marks & LaRosa, 2012),

meaning that reliability coefficients with their associated sources of construct-irrelevant variance are treated as interchangeable.

The methods used for evaluating the accuracy of classification of developmental screening tests have been criticized for introducing bias that lead to the inability to replicate the results from one study in other contexts (Sonnander, 2000; Camp, 2006; Camp, 2007). It has been noted that while certain clinical utility measures (sensitivity, specificity) are thought to be properties of the scale, they are still dependent on the naturally occurring rate of delay in the local population. This outcome leads to a high proportion of over-referrals and a low rate of true positives (Santor, 2005; Camp, 2006). Because a "gold standard" does not exist for these kinds of measures (Sonnander, 2000) and other general screening tools are used as reference tools in diagnostic accuracy studies, the co-positivity and co-negativity can sometimes be explained by different norms, diverging content and their own sources of measurement error (Camp, 2006; Camp, 2007). Finally, conditional probabilities are often calculated on small samples selected only from children with positive results on the screening instrument under review which inflates sensitivity values (Camp, 2007).

Use of non-parametric item response theory (NIRT) for item and subscale score analysis. The methods described above apply to the domain- and test-level scores, while only a few papers mention the importance of looking at the properties of individual items (e.g., Glover & Albers, 2007; Aylward & Stancin, 2008; Christ & Nelson, 2014). This is an important gap in the research literature given that item difficulty is approximated by developmental quotient ratio of mental to chronological age. Hence, interpolation and extrapolation are often used to measure average performance across age groups (Salvia & Ysseldyke, 2009). Item response theory modeling can be used to estimate item difficulty

and establish item discrimination conditioned on the level of developmental ability, as well as select the items that provide the maximal precision around the selected cut-off points placed on the ability continuum (Embretson & Reise, 2000; Livingston, 2006). For example, both classical and IRT-based methods were used to set the cut-off scores on the companion screening tool for social competencies, ASQ-SE ("Social-Emotional", see Yovanoff & Squires. 2006).

Focus on item-level results warrants special attention for the ASQ-18 because each of the domain scales is short, every item is an important task linked to development, and the domain score, separated by cut-off points, is an unweighted sum of the items. As a result, selecting this method of scoring implies that selected milestones are equally informative across all levels of the development continuum and that weights assigned to response options accurately reflects the degree of skill acquisition across populations (Santor, 2005). Applying IRT modelling to mental health screening, Santor and colleagues
(Santor, Ramsay, & Zuroff, 1994; Santor & Coyne, 2001; Santor, 2005; Santor, Ascher-Svanum, Lindenmayer, & Obenchain, 2007) demonstrated how to verify these assumptions if a trait continuum is assumed, how to assemble a shorter test from well-functioning items with the same amount of diagnostic accuracy and how to evaluate cut-off functioning if the "true" diagnosis is known. However, for a shorter depression scale, they observed that misclassifications often occur even with highly discriminating items.

Parametric IRT models have been used extensively in educational and cognitive testing for item analysis (cf. Yen & Fitzpatrick, 2006). However, their applicability to testing for pathology and clinical symptoms has been questioned because of relying on small convenience samples, mixing clinical and non-clinical samples, using poorly-defined

content domains, specifying narrow or broad constructs, and focusing on extreme regions of the ability continuum where the parametric models are traditionally less precise (Reise & Waller, 2009). Non-parametric IRT models, such as non-parametric kernel smoothing regression (Ramsay, 1991), can overcome some of these limitations, as assumptions include neither a unidimensional latent trait underlying the responses nor the specific form of the curve linking two variables together. The detailed overview of strengths and weaknesses of the NIRT, as well as the rationale for its selection, can be found in the methods section.

Ages and Stages Questionnaires: Description and Psychometric Properties

The Ages and Stages Questionnaires (ASQ) is a developmental monitoring and screening questionnaire system consisting of 21 age-specific scales. It is used to evaluate child development at selected-age intervals over the first three years of life, and to identify children at risk for developmental delays. Each of the scales contains items measuring all general domains of development, including communication, gross motor skills, fine motor skills, cognitive (termed "Problem-solving"), and personal-social development. Items represent everyday tasks associated with significant developmental milestones that are easy for parents to elicit and observe in a variety of settings (Squires et al., 2009). For all questionnaires, item difficulty was determined by the developmental quotient ratio of age equivalent for the item to the ASQ age interval, multiplied by 100. Items are expected to be of low, low-medium, and medium difficulty. In addition to scaled items, each questionnaire contains a set of general questions about the child's health to elicit parents' concerns and augment information from the scales.

ASQ-18 domain content. The specific questionnaire under consideration in this study, the ASQ-18, contains 30 items aimed at children between 17 months and 18 months 30 days should be able to perform. Below is the interpretation of tasks in the light of skills and actions expected of children at this age.

Communication. The six items in this subscale measure a child's receptive communication skills (comprehension of simple one-step commands in a familiar context and use of symbolic gestures) and productive language skills (use of words and two-word utterances). Receptive tasks correspond to developmental milestones expected to appear between 12 and 18 months (Feldman & Messick, 2007; Papalia, Olds, Feldman, & Kruk, 2008). Language production tasks refer to a progression of vocabulary acquisition that is characterised by large intra-individual variance: while the first simple words should be appearing by 18 months, vocabulary growth and telegraph-like sentences may take until the end of the 2nd year to develop (Feldman & Messick, 2007).

Gross motor and fine motor. The six gross motor items measure a child's ability to walk freely and produce other coordinated movements that require a sense of balance (climbing, kicking), thus covering activities and body functions (balance and coordination needed for activities). The six fine motor items refer to muscular coordination and controlled release (throwing a small object by moving a forward arm as opposed to dropping it) and well-developed grasping skills (stacking small objects one on another, holding a pencil or pen, turning the pages of a book and using a spoon). Most of those skills are expected to be developed by 18 months (Eliot, 1999). On average, both gross and fine motor skill development follows a set of predictable milestones common for all infants,

although the individual variability in skill acquisition may span as much as 9 months (Papalia et al., 2008).

Personal-social. The six items in this domain measure pro-social behaviour, self-awareness, ability to recognize goal-directed action and selectively imitate a predictable sequence of causal routine behaviours (Ross, Vickar, & Perlman, 2010; Metzloff & Williamson, 2010). Most items tap into several areas of development at once. For example, offering a doll to one's own image in the mirror while playing with a doll can be interpreted as showing empathy, imitating a directed action, recognizing similarity between one and others and directing joint attention to a third object (Ross et al., 2002; Rochat, 2010). Goal-directed behaviour, joint attention and imitation are expected to be formed by the age of 14 months. Between 15 to 18 months children should be able to direct attention to their needs with gestures and language and show empathy to others, thus, recognizing them as subjects. However, identifying oneself in the mirror generally happens towards the end of the 2nd year.

Problem-solving. As it is the case with the personal-social items, tasks in the cognitive domain are multidimensional and require an integration of skills and competencies (Aylward, 1997). These tasks require children to imitate goal-directed actions and reproduce them from memory after a delay, discriminate one object within another, perceive size and shape constancy, make a mental picture of an action and its consequence, connect perceptions together and translate them into fine motor movements (Hetherington et.al., 2006). Skills required for most of the items – grouping objects, identifying goal-directed behaviour and using it for the problem-solving purposes, even after a delay – should emerge by the age of 18 months. Ability to scribble without imitating a certain

shape also appears around this age, but reproducing a line is a more difficult task requiring symbolic thinking. Hence, this type of behaviour is expected to appear around 24 months of age (Hetherington et al., 2006).

Administration and scoring. According to instructions in the manual, parents can complete the questionnaire at home or in a primary care office setting, but they should ensure that the child is well-rested and ready to play, and should give the child an opportunity to try out the actions before recording the results. However, reports from the primary care setting (Rydz et al., 2006; Simard et al., 2012) indicate that parents tend to complete the tool from memory while waiting for appointments, especially when a child refuses to cooperate, a common behavior in toddlers exposed to new actions or environments (Berger at al., 2010). Even with trying out the items, the questionnaire takes about 15 minutes to complete.

Each item is scored using a three-point scale: 0 for "no", 5 for "somewhat" and 10 for "yes", and then summed separately for each domain. On the latest revision of the ASQ scales, children are categorized into three groups based on their score: "at-risk" (2 standard deviations below the mean), "monitoring zone" (between 2 and 1 SD below the mean) and "typical" development (the rest of the distribution). At-risk children should be immediately referred for further diagnostic assessment or intervention while the children falling into the "monitoring zone" should be given specific follow-up activities to improve skills in need of intervention and should be regularly re-screened. Although the statistical (distribution-dependent) definition of delay was used for this scale, Canadian practitioners rely on American cut-off scores for interpretation. However, distributional methods of setting cut-

off scores are common in health measurement in general (Streiner & Norman, 2008) and developmental screening in particular (Aylward, 2009).

Psychometric properties. The norms for the current version of the ASQ were developed on the US national sample of children. The sample size for the 18-month questionnaire was 616. Thirty percent of children had one or more biological or environmental risks and children were recruited partly from the early intervention programs, meaning that those children were likely experiencing delays. Reported measures of psychometric quality included the following:

1. Test-retest reliability analysis was conducted on questionnaires filled by 145 parents at two weeks interval. The results showed 92% agreement and inter-rater reliability between parents and trained examiners. Both were measured by intra-class correlations for some periods, which were not identified further in the manual. The test-retest ICC range was 0.75-0.92; the interrater agreement was 93% agreement, with the ICC range 0.43-0.69.
2. For 12-month and 24-month questionnaires only, the partial credit model was use to investgate presense of differential item functioning (DIF: Camilli, 2006) in paper vs. web-completed questionnaires. DIF was found on only a small number of items. Model-item fit was not reported (Yovanoff, Squires, & McManus, 2008).
3. Internal consistency was tested for all age intervals using Cronbach's coefficient alpha. For the 18-month questionnaire, coefficient alpha ranged from 0.54 to 0.58 for three domains (fine motor, problem-solving and personal-social) and was higher for communication (0.74) and gross motor skills (0.77).

4. To set the cut-off scores, a subset of children with known status (both at-risk and not at-risk) were administered both ASQ and another general developmental screening tool, Batelle Developmental Inventory (Newborg, Stock, Wnek, Guidubaldi & Svinicki, 2004). Both second and third editions of the ASQ questionnaires were administered. Cut-off scores were selected to maximize both sensitivity and specificity using receiver operator characteristic curves on the conditional probabilities. The authors addressed many of the common limitations, including selecting both at-risk and typically developing children and making data for all specific intervals available. However, only 22 children in total, 11 of them with an at-risk result on one of the ASQ domains, were in the 18-month group. Sensitivty and specificity values were in the acceptable range (i.e., above 0.7), but 18% of children were overidentified by ASQ confirming the problems with low positive predictive value noted by many researchers (Camp, 2006; Glascoe & Dworkin, 2008; Pool & Hourcade, 2011).

Although the ASQ is recommended for use with any general pediatric and at-risk population (Drotar et al., 2008), concerns have been raised about its accuracy in some contexts, including its use with children at low birthweight (Schonhaut, Armijo, Schonstedt, Alvarez, & Cordero, 2013), premature children (Aylward, 2005; Rydz et al., 2006; Simard et al., 2012) as well as with multicultural populations (Gokiert et al., 2010). As far as the use of US-based norms in Canada is concerned, several studies that have been conducted on the use of U.S. norms with Canadian children found statistically significant difference between cut-off locations in Canadian samples compared to the US normative sample (Dionne, Squires, Leclerc, Peloquin, & McKinnon, 2006; Rydz et al., 2006; Simard

et al., 2008; Frisk et al., 2009; Limbos and Joyce, 2011). A study by Rydz et al. (2006), conducted in a similar context – screening of 18-month children in a general pediatric practice in Quebec – raised concerns about the high rate of false positive results and highlighted the fact that children's status can change within three months of administering the tool. However, these studies were conducted with the previous version of the ASQ, with a different set of cut-off scores. The second version also had only one cut-off point, separating at-risk from typically developing children, with no monitoring zone in between. This limits the transferability of these results to the current version of the tool.

Statement of the Problem and Research Question

Analyses using data from the ASQ-18 have shown good psychometric properties in the US normative sample (Squires at al., 2009), but the methods used in validation may be sample dependent and the cut-off scores may only be applicable to US children. The norming sample was representative of the US population but contained a higher proportion of children with known risk factors than would be expected for a general pediatric population in primary care (30% of children with at least one known risk, vs. 1 to 10% cited above).

As the ASQ-18 is still used with children at regular (requiring no referral) visits to providers of health and family services in Alberta, the primary purpose of the study is to evaluate the psychometric properties of the tool, including the comparison of the cut-off location, in the convenience sample from a local primary care population in Alberta and to make inferences about the potential consistency of classification with this tool in the Canadian context.

The secondary purpose is to explore the use of alternative psychometric methods and models, particularly when a criterion variable or reference standard is not available. Assuming the continuum of development, item and domain subscale functioning will be examined at the low to medium developmental levels and precision of the scale around the cut-off scores will be estimated using a variety of psychometric methods. The results will be logically linked to how well the ASQ-18 can differentiate typically developing children from those who need immediate secondary assessment or monitoring. Thus, the research question is as follows:

Does the ASQ-18 with American norms have appropriate psychometric properties to consistently and precisely discriminate between children at-risk, in need for monitoring, and typically developing, in the Canadian primary care context?

Chapter 3. Methodology

This chapter begins with a description of the data used in the study, focusing specifically on the difference between the Canadian and US normative data. It continues with a brief overview of all methods used to obtain domain-level evidence of reliability and precision of scores as well as evidence of item functioning at the low ability levels. It also provides a detailed description of the methods including: (a) the use of polychoric correlations to measure internal consistency, (b) conditional standard error of measurement for observed scores associated with the cut-off scores, and (c) non-parametric Testgraf. Taken together, these methods provide information about the tool's capacity to discriminate among children at various developmental ability levels in a convenience sample from a general pediatric population in Alberta.

Data

The data used in this study were collected by a Western Canadian health authority between 2007 and 2010 as a part of a pilot screening project. The questionnaires were completed by parents during immunization appointments at clinics and community organizations in a large Canadian urban centre. The database contained basic information about the children (age, gender, main language, cultural background) and their families (country of origin, income, guardians who filled in the questionnaire, type of childcare). The responses from the questionnaires where parents needed the help of translators were not included in the analysis to limit construct-irrelevant variance attributable to cultural differences in child-rearing (Gokiert et al., 2010). The final sample contained 1,009 cases.

As the data were collected in the community for purposes other than validation, they have several limitations that render many of the original psychometric analyses impossible to replicate.

1. No secondary screen or full-scale assessment was available for the majority of children. This ruled out the "gold standard" methods of cut-off score validation (e.g., the item response characteristic curve method for establishing original cut-off scores).
2. The screening tool was completed only once, in one setting (clinic), and mostly by mothers. Thus, the influence of time, setting and rater effect on reliability and decision consistency could not be estimated.
3. No specific attempt was made to recruit more at-risk children into the sample. As a result, most children easily performed the activities described in the scales and the distributions of both individual items and the summed domain scores tend to be negatively skewed.
4. Unlike the distribution of the US sample, the Canadian data were not normed and the cut-off scores cannot be expected to match percentiles of the normal distribution. In addition, both domain scores and item scores behaved like discrete variables in preliminary analyses (i.e., scores were separated by clear intervals of 5 points) diminishing the accuracy of correlation-based methods such as Cronbach's alpha.

In the light of limitations, the present study will focus on the following aspects of cut-off score functioning in a new population

Overview of methods

This section provides a general overview of methods, used in this study. The next section follows with an in-depth description of selected methods of subscale and item analysis that are not common in the context of developmental screening.

1. Domain score reliability and the standard error of measurement (SEM). This information was obtained using correlational CTT methods. Polychoric correlations are used for internal consistency estimates, because they provide a better fit to the scale properties, compared to Pearson correlations. The average scale SEM was complemented by the conditional SEM (CSEM) estimates under the multinomial error model (Lee, 2005).
2. Preciseness of the domain scores at the target levels of development or ability. This information comes from the application of the non-parametric item response theory (NIRT) implemented in Testgraf (Ramsay, 2001).
3. Difficulty and discrimination of single items within domains to explain evidence from the domain scores. This information comes from option and item characteristic curves estimated in Testgraf as well as from item-rest score correlations, i.e. correlations between an item and the domain score with the item removed (Ramsay, 2001).

Description of the Selected Methods

Polychoric correlation. A polychoric correlation is a measure of the relationship between two ordinal or discrete manifestations of underlying continuous normally distributed latent variables (Hershberger, 2005). The ASQ-18 item and domain scores can be assumed to represent an underlying continuum of development but on a

restricted measurement scale. The purpose of a polychoric correlation is to correct for this restriction of range. Polychoric correlation defines a variable *Y* with c_1, c2…c_n categories as a latent continuous variable *Y'* with *C-1* thresholds estimated from observed categories:

$$Y = c, \tau_c < Y' \leq \tau_{c+1,} \; where \; \tau_{0=} \; -\infty, \tau_{c=} \; +\infty$$

Assuming the bivariate normality, probability of a response Y falling into a category c is

$$p(Y = c) = \; p(Y' \leq \tau_{c+1)} \; - p(\, Y' \leq \tau_{c)} \,,$$

where $(\, Y' \leq \tau_{c)} \; = \Phi(\frac{\tau_c - \mu}{\sigma}\,)$, Φ is the cumulative normal probability density function and $\mu \; and \; \sigma$ are mean and standard deviation of the latent variable Y'. Probabilities of falling within each possible category are summed and the correlation between the underlying variables is estimated using limited information maximum likelihood (Hershberger, 2005). The polychoric correlation matrix, estimated in PRELIS (Jöreskog & Sörbom, 1996), was used in this study to obtain item- rest score correlations and to obtain the ordinal version of Cronbach's alpha, as described by Zumbo, Gadermann, and Zeisser (2007). A simulation study by these authors suggested that the ordinal alpha provided a more accurate estimate of reliability under the conditions of skewed response distribution, small number of response options and lower internal consistency due to the diversity of item content.

Conditional standard error of measurement. According to the Standards for Educational and Psychological Testing (AERA, APA & NCME, 1999), an average measure of scale consistency, such as Cronbach's alpha, needs to be complemented by the conditional standard error of measurement (CSEM) when the scale is used to make classification decisions (Standard 2.14, p.35). This information is needed to more accurately estimate the error around the cut-off scores. Given the polytomous structure of the ASQ, the multinomial error model was deemed appropriate. This model was proposed

by Lee (2005) as an extension of Lord's binomial error model used with dichotomous items (Lord, 1965, cited in Nunnally & Bernstein,1994, p. 242-243). The model represents a total domain score for each child, denoted *Y,* as $Y=c_1X_{1+}\ c_2X_{2+\ldots+}\ c_kX_{k,,}$ where $c_{1,}c_{2\ldots}\ c_k$ are possible scores on an item, and $X_{1,}\ X_{2\ldots}X_k$ is a random variable signifying the number of items scored as *c*. X follows a multinomial distribution defined as

$f(x1, x2 \ldots xk) = \frac{n!}{x1!x2!\ldots xk!}\ \pi_1^{x1}\pi_2^{x2} \ldots \pi_k^{xk}, x \in \Omega$, where Ω is the space of X.

The assumption of randomly parallel tests means that the same score *c* will be given to a child for any proportion of parallel items π sampled from the item universe provided that the number of items and number of item score points are identical. In this model, the unbiased estimate of CSEM for a child with proportion score $\pi_i\ = \frac{x_i}{n}$ is defined as

$$\hat{\sigma}_{e(Y)=}\ \hat{\sigma}_Y = (\frac{1}{n-1}\left[\sum_{i=1}^{k} c_1^2\ x_i(n-\ x_i) - 2\sum_i\ \sum_j\ c_i\ c_j x_i x_j\right])^{-\frac{1}{2}}\ ,$$

where $2\sum_i\ \sum_j\ c_i\ c_j x_i x_j$ is the covariance of errors. In calculation, π_i is replaced by observed mean proportion score, $\overline{x_p}$ as it is an unbiased estimator.

Non-parametric IRT/Testgraf. The observed score subscale -level statistics in this study were complemented by item and test information analyses conducted using the non-parametric item response theory framework. The Testgraf procedure (Ramsay 1991, 1997, 2001) selected for this study is one of the few NIRT techniques that can be used with a polytomous scale. The estimation of item and ability parameters works as follows:

1. The observed domain score for each child is transformed into a quintile of the standard normal distribution, which serves as the initial estimate of ability, $\theta_{1,}$ $\theta_{2\ldots}\ \theta_{a,}$ where a=N+1.

2. For each response option *m* of item *I,* conditional binary response vector y_{ima} is calculated.
3. To speed up further calculations, the number of evaluation points at the ability scale is reduced to Q=51 equally spaced intervals (51 is the default option used in the analysis).
4. The area of the standard normal curve φ_r is computed for each rank *r* falling into the intervals adjacent to each evaluation point q. These rank values, θ_r, become the new ability values.
5. Values of y_{ima} conditioned on θ_r are transformed via weighted averaging using the transformed into response vectors, $y_{imr} = \hat{p}_{imr}$ using Nadaraya-Watson kernel function. This function assigns weights based on the proximity of θ_r to θ_q, the evaluation point, where larger distance from the evaluation point leads to smaller weights.

The final formula for the option response pattern m of item i for each θ_q is as follows:

$$P_{im}(\theta_q) = \sum_{r=1}^{Q} w_{rq} p_{imq} = \frac{\sum_{r=1}^{Q} \varphi_r K\{\frac{\theta_r - \theta_q}{h}\} p_{imr}}{\sum_{r=1}^{Q} K(\frac{\theta_r - \theta_q}{h})},$$

where $P_{im}(\theta_q)$ is the probability of endorsing option m of item I given ability level θ_q; $\frac{\varphi_r K\{\frac{\theta_r - \theta_q}{h}\} p_{imr}}{K(\frac{\theta_r - \theta_q}{h}}$ is the kernel function that uses Gaussian kernel $K(u) = \exp(-u^2/2)$ and bandwidth parameter h to determine local averaging weights, w_{rq} for each conditional option response vector p_{imq}.

The option response functions are then summed up at each ability level to produce item response functions and the test (domain) information function. The test information function is a parametric estimate that uses logistic-quadratic function and is defined as a sum of item information functions

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta), \text{ where } I_i(\theta) = \sum_{m=1}^{M} \left(\frac{dPim}{d\theta}\right)^2 / Pim(\theta).$$

The standard error of measurement is also taken to be the inverse of the square root of the TIF.

The flexibility of local averaging leads to a more accurate modelling of item-domain score relationship in the extreme areas of the score scale when compared with parametric models, because of lower sample sizes in these score regions. This method does not impose a certain distribution on the response probabilities thereby making it easier to evaluate item and option discrimination at various ability levels. In turn, evaluating effectiveness of items at different levels of ability can be used to improve the capability of the scale to discriminate between children at ability levels falling into different categories.

Chapter 4: Results

This chapter is focused on the results of the study. It includes a description of the psychometric properties of ASQ-18 with the goal of evaluating whether the scale allows for accurate identification of children who may require developmental intervention or follow-up monitoring. The results will focus on (a) a description of the sample and observed domain scores; (b) location of cut-off scores in the Canadian score distribution compared to the US normative sample distribution; (c) side-by-side comparison of cut-off locations in the observed score distribution and the ability score distribution estimated by NIRT; (d) evidence for the reliability and precision of the domain subscale scores[1] around the cut-off scores; and (e) the psychometric quality of individual items as determined by difficulty and discrimination for the Canadian sample.

Sample and Data Source

Data used in this study were collected from children between 17 and 19 months of age whose families participated in the developmental screening pilot in a multicultural community in a large city in Alberta. Out of the initial sample of 1,807 children, 1,009 were used for the analysis. The remaining 798 cases were deleted due to incorrect age questionnaire being administered, premature status, or the need for cultural brokering. The latter two groups were removed from the analysis as previous research has shown mixed results regarding the validity of ASQ results for preterm infants (Simard et al., 2012) as well as culturally diverse children (Gokiert et al., 2010). All data were collected in a primary care setting while parents and their children were waiting for the immunization appointments at a public health centre. Overall, about 80% of the children in this

[1] The terms "domain" and "subscale" will be used interchangeably in this study.

community were served by this appointment over the course of the pilot. Thus, the sample used in this study can be considered representative of the general pediatric English-speaking population of this community.

This Canadian sample is comparable to the US normative sample on some demographic variables, but also has several important differences. Both samples have a similar proportion of males and females (49% and 51% in the Canadian sample, respectively, and 50% each in the US sample). In both demographic samples the primary caregivers who filled in the questionnaire were primarily mothers (81 and 82%, respectively). The majority of the families in the Canadian sample listed English as their main language (70%) or secondary language (25%) used at home, and 61% identified Canada as their country of origin. Thirty-five percent of the respondents reported that their annual income was higher than the province's median income for large urban centres. Another 25% reported the household income to be in the range below the median. SES and ethnicity data cannot be meaningfully compared to the US sample because these categories are defined differently in the two countries.

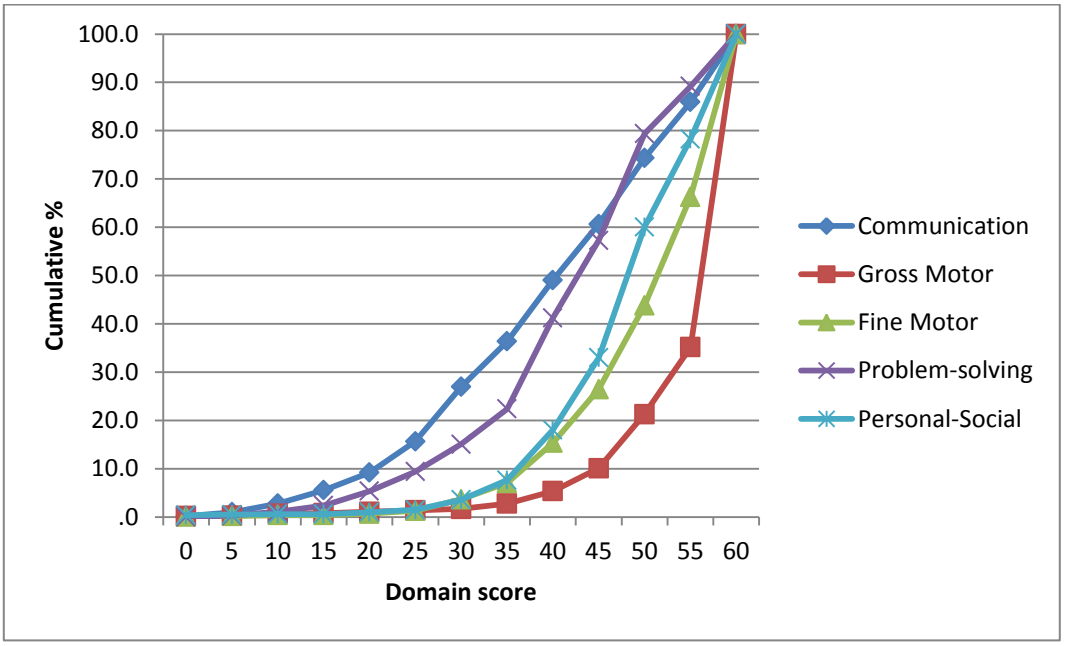
Two important differences exist between the two samples. First, the American normative data were gathered at preschool developmental and remedial programs and had a higher proportion of children at-risk for developmental delays. As the Canadian screening program targeted all children in the community, regardless of their health status, the proportion of children at risk for delays is expected to be lower. Second, while the American parents completed print and online version of the questionnaire both at home and in the clinics, the Canadian parents in the sample only completed the questionnaire at the clinic. They possibly had limited completion time and filled it from memory without

observing the child doing the required actions. Such implementation can negatively impact the accuracy of screening results (Marks, 2007).

Domain-Level Evidence

Descriptive statistics. Figure 1 presents the cumulative percentage distribution of domain scores for all subscales. As could be expected from a sample coming from routine primary care visits, most children fall into the "healthy development" zone according to their score making the distribution of all domain scores positively skewed. The proportion of children falling into the at-risk category does not exceed 3.5% for four out of five domains, with the exception of 9.4% in the problem-solving domain. The proportion of children falling into the monitoring zone ranges from 7.6% to 27% across all domains. These results are consistent with the distribution of scores on similar tools for children reported in other studies of ASQ (e.g. Rydz et al., 2006). Gross motor, fine motor and personal-social domains appear notably less difficult than communication and problem-solving domains.

Figure 1. Cumulative distribution of domain scores.



Cut-off score location in Canadian and US samples. Table 1 illustrates how the Canadian sample compares to the American sample in terms of summary statistics (mean, standard deviation) and in the location of the cut-off scores, expressed in z-score units and percentiles of the standard normal distribution. The discrepancies between the two distributions are in bold font. In this table and in the subsequent tables, label "c2" stands for the cut-off score separating the at-risk category from the monitoring zone category, while "c1" stands for the cut-off score between the monitoring zone and the healthy development category.

Table 1. Comparison of cut-off location in Canadian and normative US sample

	Canadian sample					Normative sample				
	M(SD)*	Z(c2)**	P(z≤c2)	Z(c1)**	P(z≤c1)	M(SD)*	Z(c2)	P(z≤c2)	Z(c1)	P(z≤c1)
Communication	41.62 (13.63)	-2.10	0.02	-0.85	0.20	42.30 (14.62)	-2.00	0.02	-0.84	0.20
Gross Motor	55.96 (7.70)	**-2.41**	**0.007**	**-1.24**	**0.11**	55.46 (9.04)	**-2.00**	**0.02**	**-1.00**	**0.16**
Fine Motor	51.71 (8.87)	-1.96	0.02	-0.92	0.18	52.44 (9.06)	-2.00	0.02	-0.99	0.16
Problem-Solving	43.85 (11.14)	**-1.62**	**0.05**	**-0.72**	**0.237**	45.99 (10.13)	**-1.99**	**0.02**	**-1**	**0.16**
Personal-Social	49.74 (8.84)	**-2.55**	**0.005**	**-1.38**	**0.08**	**47.90 (10.35)**	**-2.00**	**0.02**	**-1.00**	**0.16**

The comparison shows that while the average score and variability are close in both distributions, the proportions falling below and above either cut-off score differ in three out of the five domains. In the normative sample, the relative position of the cut-off scores corresponds to the position of -2 SD and -1 SD from the mean in a standard normal distribution, with the exception of the monitoring zone cut-off point for the communication domain (20^{th} percentile as opposed to the 16^{th} percentile). In the Canadian sample, cut-off location for the communication and fine motor domain matches the location in the normative sample; it has moved closer to the mean for gross motor and personal-social domains and closer to the mean in the problem-solving domain.

Cut-off location differences can result in potential misclassification of children on the affected subscales. Aligning the cut-off scores with 1 and 2 standard deviations below the mean in a Canadian distribution would result in different observed scores being associated with the category boundaries. The observed scores falling immediately below the cut-off scores would be 5 points higher in the gross motor, personal-social domain and problem-solving domains. While only 2 to 3% more children would be falling into the at-risk group, if the cut-offs were adjusted, the percentage of children falling into the monitoring zone in this case would increase by 10%. In the problem-solving domain, the percentage of children falling into these categories would decrease by 5% and 7%, respectively. These preliminary estimates suggest that the current cut-off location can potentially lead to over-identification in the latter and under-identification in the former domains.

Cut-off location in observed and ability score metric. Table 2 shows location of the cut-off scores in the observed distribution and in the ability distribution estimated by Testgraf and expressed in standard score and expected score metrics.

Table 2. Cut-off score equivalents in the observed and NIRT ability distributions

	Cut-off point	Exact	Observed Score*	Observed percentile	Normal Theta	Expected theta
	C2	13.06	10	2.8	-2.64	2.03
Communication	C1	27.68	30	**27.0**	-1.08	5.85
	C2	37.38	35	2.8	-	-
Gross motor**	C1	46.42	45	10.1	-	-
	C2	34.32	30	3.7	-2.76	5.94
Fine motor	C1	43.48	40	**15.4**	-1.56	8.35
	C2	25.74	25	**9.4**	-1.92	4.61
Problem Solving	C1	35.86	35	22.4	-1.08	6.85
	C2	27.19	25	1.5	-2.52	5.59
Personal-Social	C1	37.55	35	7.6	-1.92	7.20

* max=60 ** NIRT estimates could not be obtained for the gross motor domain due to lack of monotonicity

The observed scores differ from the exact cut-off scores, as the distribution is discrete. For the most part, the scores immediately below the exact cut-off mark the boundaries of a category. The communication subscale is an exception, as the score 27.56 has been rounded up to 30 by practitioners as per test manual instructions. The observed scores corresponding to exact cut-off scores will be henceforth referred to as "cut-off" scores because the exact scores do not appear in the discrete Canadian distribution. In the case of the NIRT estimates, several ability scores could correspond to a particular raw score point due to local averaging. To address this problem, the most representative score on the normal ability scale, i.e., the score that was closest to the average and most frequently matching the given observed score was selected as the location

equivalent. Expected score, which is originally derived from standard normal quintiles, comes from the Testgraf output. This table will serve as a reference point for the future analyses, especially when several methods of estimation are compared.

Table 2 also lists percentile points for the cut-off scores in the observed distribution. These points can be compared with the normal percentile equivalents in Table 1. For most domains, the difference between the observed and standard normal percentile does not exceed 3 percentile points (bolded numbers indicate a larger difference). According to the review of raw percentile distribution for the affected domains, this discrepancy resulted in different scores on four out of the total of 18 items (two items in the communication domain, one item each in problem-solving and gross motor domains). This finding suggests that a relatively good fit of the kernel-smoothing function based on the Gaussian kernel can be expected at the lower ends of the ability scale and allows for the proper use of the standard normal distribution percentiles in future analyses based on the observed scores. However, standard normal quintiles, produced by Testgraf, are less suitable for direct comparison with normal distribution percentiles reported in Table 1, as the values differ by 0.3 to 0.8 standard units in three out of four domains. Thus, these results will be reported separately.

Reliability and Precision. This subsection explains the findings in Table 3, which pertain to the evidence for the reliability properties of the ASQ-18 scores, which are pre-requisites for the tool's capacity to consistently screen children in the current sample. First, the internal consistency index (i.e., Cronbach's alpha)

was calculated using polychoric inter-item correlations to compare the findings to those reported for the US normative sample. Alpha results were also used to calculate 95% confidence intervals with the relative standard error of measurement (SEM). This measure of SEM, or the average standard error for all domain scores, was complemented by the estimate of the conditional standard error of measurement (CSEM) for the observed scores. To relate the amount error to the ability levels, NIRT test information function, and CSEM of the ability estimates were calculated. The latter were used to construct 95% confidence intervals around the ability levels corresponding to the observed cut-offs.

Internal consistency. Internal consistency estimates show acceptable level of reliability for all domains, but the standard error of estimate may be large enough to make misclassification possible. Cronbach's alpha estimates for the subscales ranged from 0.69 to 0.89, which can be considered acceptable given the heterogeneity of items (Gadermann, Guhn, & Zumbo, 2012). Alpha is higher for more homogeneous domains: communication and gross motor and meets the standard recommended for applied low-stakes settings (0.8;Gadermann et al., 2012). For other domains, the lower alpha level can be related to potential heterogeneity of items, also reflected in the medium strength of inter-item correlations and weak minimum possible correlations between some items.

Table 3. Reliability and error of measurement for ASQ-18

	Communication	Gross Motor	Fine Motor	Problem-solving	Personal-Social
C2 (at-risk)	13.06	37.38	34.32	25.74	27.19
C1 (monitoring)	27.68	46.42	43.48	35.86	35.77
Internal consistency reliability					
$\bar{r}_{\mathbf{i}}$ (polychoric)	0.40	0.57	0.28	0.30	0.28
$\boldsymbol{r_{min\text{-}max}}$ (polychoric)	0.05-0.72	0.30-0.86	0.11-0.73	0.02-0.83	0.07-0.46
Alpha (polychoric)	0.83	0.89	0.69	0.72	0.69
95% CI around					
At-risk cut-off	(2.05,24.08)	(32.37,42.39)	(24.64,44.00)	(14.20,37.28)	(17.55,36.83)
Monitoring cut-off	(18.99,41.01)	(41.41,51.42)	(33.80,53.16)	(24.31,47.40)	(27.91,47.19)
CSEM					
At-risk cut-off					
$\bar{\sigma}_{e(y)}$	7.96	10.34	11.51	10.89	9.22
$\sigma_e(min - \max)$	6.32-10.00	6.71-10.25	7.75-13.42	5.00-12.04	5.00-12.04
Monitoring zone cut-off score					
$\bar{\sigma}_{e(y)}$	12.59	10.00	10.25	11.06	11.29
$\sigma_e(min - \max)$	0-13.42	9.22-12.04	6.32-12.04	5.00-12.04	5.00-12.04

NB. $\bar{r}_{\mathbf{i}}$ – average inter-item correlation; $\bar{\sigma}_{e(y)}$ – average CSEM around the given cut-off

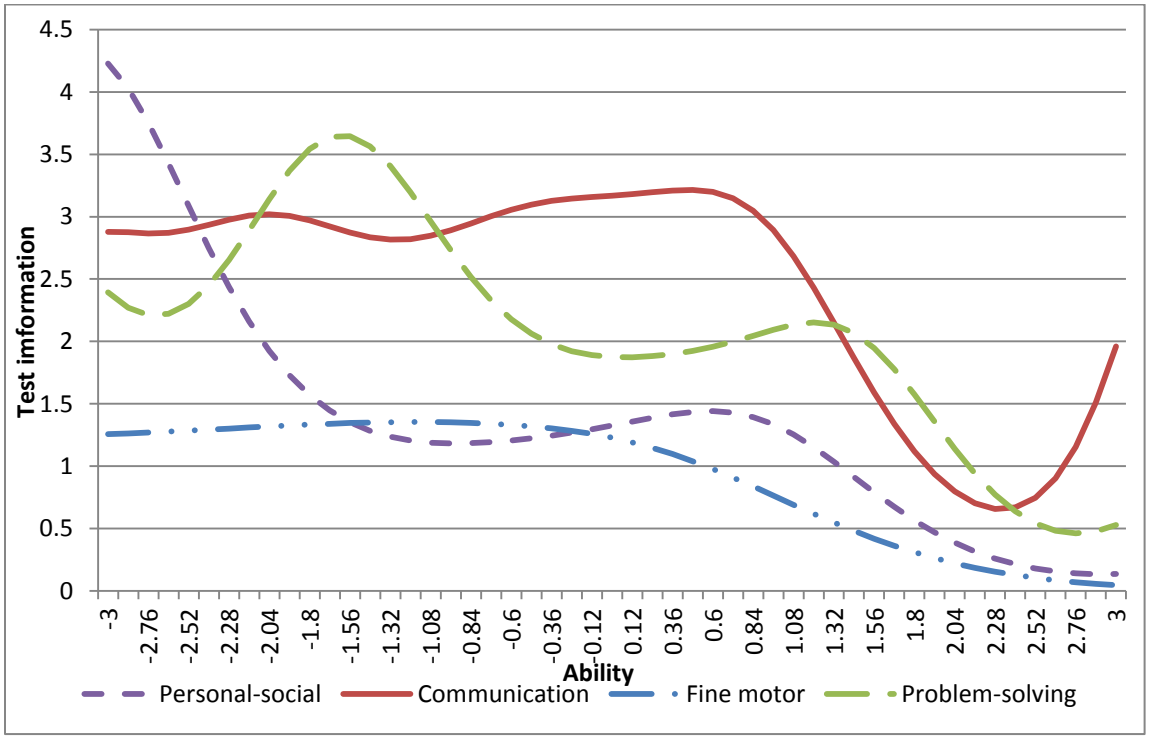
Gaderman, Guhn, and Zumbo (2012) suggest interpreting the magnitude of polychoric correlations using Cohen's effect size guidelines (Cohen, 1988, cited in Gaderman, Guhn, & Zumbo, 2012, p.5)**.** Following these guidelines, mean inter-item correlations for the gross-motor subscale can be classified as high strength while inter-item correlations for other subscales are considered medium strength. The differences in the magnitude of alpha between domains are consistent with the Pearson-based alpha coefficients as reported for the normative sample (Squires at al., 2009). Those coefficients range from >0.7 for communication and gross motor domain to 0.58-0.54 for the other domains. Although the correlations in the US sample are lower due to the Pearson's r properties (effect of homogeneity of variance and restriction of range), proportional differences in magnitude between subscale reliabilities confirm heterogeneity of item content in fine motor, problem-solving and personal-social domains.

Despite the high reliability, the standard error of measurement in each domain, calculated using polychoric alpha, produces a confidence interval that reaches into the adjacent zones for each cut-off score. The interval width presents a problem for scores in the monitoring zone, as children with true scores falling into this zone also have a 95% chance to have an observed score that falls either into the at-risk zone or the typical development zone.

Conditional standard error of measurement .Large CSEM values around the cut-off scores, the averages of which are mostly larger than the distance between the observed cut-offs, confirm the possibility of misclassification. CSEM values present a more accurate estimate of errors specifically around the cut-off scores, although they are expected to be larger than relative errors linked to internal consistency reliability (Lee, 2005). Still, the magnitude of CSEM points towards potential lack of classification consistency and under referral if child's true state is not determined by other methods (e.g., additional questions on the scale, clinical judgment).

Test information function around the cut-offs. Figure 2 compares the amount of information produced by the items in each domain subscale at various points on the ability scale. The measurement scale on the y-axis are test information units for the standard normal ability estimates. These estimates are only provided for four domains that could be analysed with Testgraf.

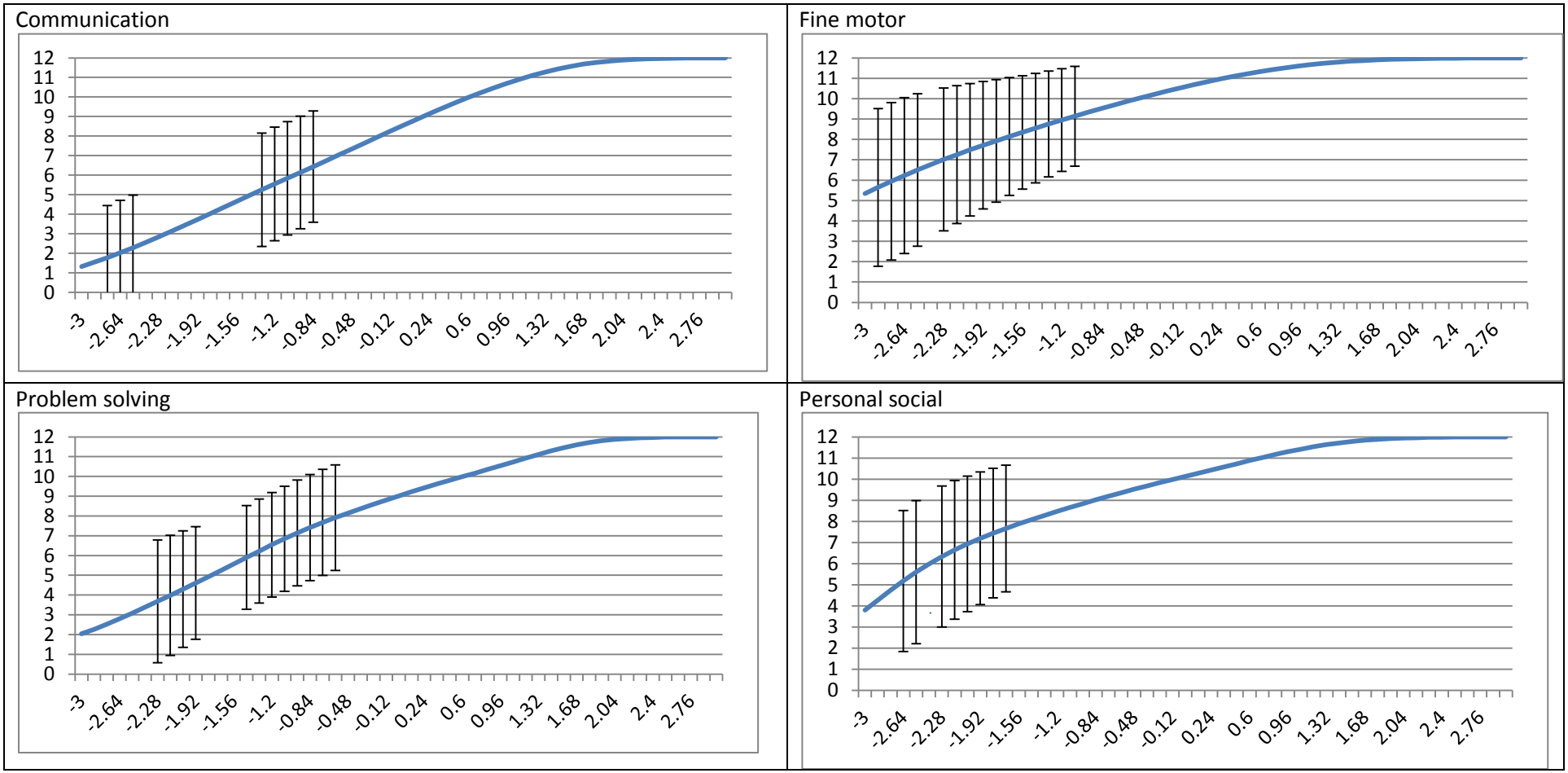
Figure 2. Test information functions for ASQ-18.



All four domains provide their maximum information around the ability levels corresponding to the at-risk cut-off point, but the maximum possible amount of information is different across domains. The problem-solving and personal-social domain curves peak around their respective at-risk cut-off scores (ability scores of -1.92 and -2.52, respectively), but steeply decrease around the monitoring zone cut-off scores (ability scores of -1.08 and -1.92, respectively). The fine motor domain provides the lowest amount of information, as it does not exceed 1.5 out of a possible 4.5 on the information scale. The communication domain show as much information around the monitoring zone cut-off scores as around the at-risk cut-off score (ability scores of -2.64 and -1.08, respectively), and provides a consistently high amount of information (3 out of 4.5) for all below-average to average ability levels.

Conditional standard error of the ability estimate. Standard error of the ability estimate was calculated using TIF values as an additional method for estimating conditional standard error around the cut-off score and for determining scale precision around the target ability levels. Similar to the standard error of measurement obtained from the internal consistency estimates, the NIRT estimate of SEM was used to estimate the 95% confidence interval around the cut-off scores. These intervals are displayed in Figure 3 for the expected value equivalents of the cut-off scores (see Table 1 for the list of values). As more than one observed score could correspond to the expected score, due to local averaging, confidence intervals are also displayed for the expected scores in the vicinity of scores corresponding to the observed cut-offs.

Figure 3. SEM of the ability estimate and confidence intervals in expected score ability metric*



*To get the expected scores, items were coded as "0"-No, "1"-Sometimes, "2"-Yes. For domain subscale, min=0, max=12.

The results in Figure 3 generally align with the information provided by CTT confidence intervals and indicate scale discrimination as another source of scale precision. Even with NIRT SEM providing narrower confidence intervals than the CTT estimates (Ramsay, 2001), confidence intervals in most domains still span 40% to 50% of the scale. Therefore, children with scores in the at-risk zone (possible expected score range <7 for the fine motor domain and <5.5. for other domains) have a 95% chance of obtaining true scores falling into the adjacent zone in all domains. Similarly, children with monitoring zone scores (possible expected score range between 5 and 9 for most domains) have a 95% chance of being misclassified in either direction. The probability of the exact direction of misclassification (e.g., probability of a monitoring zone score falling into the healthy development zone, which would be a major concern) is hard to determine without the criterion variable. However, SEM estimates are expected to be less accurate at the lower levels of the ability scale due to the low sample size and high smoothing factor required by Testgraf.

Another finding arising from this figure is that scale discrimination and distance between the cut-off points also contributes to scale precision. In Testgraf, discrimination can be assessed visually by comparing steepness of the curve against the ability scores. In the case of domains with less difficult items, such as personal social and fine motor, lack of discrimination compared to the domains with more difficult items seem to be associated with the confidence intervals for the at-risk cut-off reaching into the typical development zone. Another reason, however, is the imprecision of the local averaging due to the small sample size

around the at-risk cut-off. For more difficult domains, the shorter distance between cut-off scores in the case of the problem-solving domain resulted in the CI for the monitoring zone spanning both adjacent zones. At the same time, in both the communication and problem-solving domains, the CI for the monitoring zone span a third to a half of the typical development zone score range, indicating potential for underidentification. Item-level evidence provides further context to the domain-level reliability and precision estimates, as those are in part determined by the item performance in the given population.

Item-Level Evidence

This section reviews the difficulty and discrimination of ASQ-18 items. First, these properties are described for observed scores, and then NIRT estimates are used to link the item properties to the ability levels.

Descriptive statistics. Table 4 presents the distribution of scores for each item across select percentiles: quartiles, percentiles corresponding to at-risk and monitoring zone cut-offs in the Canadian sample and in the normative sample. Bolded columns highlight the difference in responses for the percentiles corresponding to the Canadian and US normative sample cut-off score locations in the observed distribution. Percentiles provide initial estimates of item difficulty in the Canadian sample while the item-rest score correlation can be used to estimate item discrimination.

Table 4. Distribution of item responses and item- total correlations.

	Item #	Percentile points*							$r_{item\text{-}rest\ scorel}$
		C2(C)	C2(US)	C1(C)	C1(US)	25	50	75	
Communication	1	5.00	5.00	10.00	10.00	10.00	10.00	10.00	.245
	2	.00	.00	10.00	10.00	10.00	10.00	10.00	.504
	3	.00	.00	.00	0.00	5.00	10.00	10.00	.625
	4	.00	.00	.00	0.00	.00	5.00	10.00	.675
	5	.00	.00	5.00	5.00	5.00	10.00	10.00	.535
	6	.00	.00	.00	.00	.00	.00	5.00	.636
Gross Motor	1	**0.00**	**5.00**	10.00	10.00	10.00	10.00	10.00	.746
	2	5.00	5.00	10.00	10.00	10.00	10.00	10.00	.748
	3	.00	.00	10.00	10.00	10.00	10.00	10.00	.555
	4	.00	.00	10.00	10.00	10.00	10.00	10.00	.579
	5	.00	.00	5.00	5.00	10.00	10.00	10.00	.576
	6	.00	.00	5.00	5.00	10.00	10.00	10.00	.447
Fine Motor	1	.00	.00	5.00	5.00	10.00	10.00	10.00	.219
	2	.00	.00	5.00	5.00	10.00	10.00	10.00	.606
	3	.00	.00	10.00	10.00	10.00	10.00	10.00	.396
	4	.00	.00	5.00	5.00	5.00	10.00	10.00	.475
	5	5.00	5.00	10.00	10.00	10.00	10.00	10.00	.341
	6	.00	.00	5.00	5.00	5.00	10.00	10.00	.280
Problem Solving**	1	5.00	5.00	10.00	10.00	10.00	10.00	10.00	.484
	2	.00	.00	.00	0.00	.00	5.00	10.00	.669
	3	.00	.00	**10.00**	**5.00**	10.00	10.00	10.00	.732
	4	.00	.00	**10.00**	**5.00**	10.00	10.00	10.00	.562
	5	.00	.00	.00	0.00	.00	.00	5.00	.619
	6	.00	.00	5.00	5.00	5.00	10.00	10.00	.749
Personal Social	1	.00	.00	0.00	.00	.00	5.00	10.00	.270
	2	.00	.00	**5.00**	**10.00**	10.00	10.00	10.00	.242
	3	.00	.00	**5.00**	**10.00**	10.00	10.00	10.00	.412
	4	.00	.00	**5.00**	**10.00**	10.00	10.00	10.00	.332
	5	.00	.00	**0.00**	**5.00**	5.00	10.00	10.00	.276
	6	.00	5.00	10.00	10.00	10.00	10.00	10.00	.384

*Options: "0"-"Not yet", "5" – Sometimes, "10" – yes.

**Cut-off location in the US distribution corresponds to the lower percentile points, than in the Canadian sample.

As each domain subscale only has six items and the correlation may be inflated by keeping an item as part of the total score, the polychoric correlation of each item with the domain rest score (domain score minus the item score) is

reported as a measure of item-total domain score correlation. Item-rest score correlations avoid overestimation of the correlation magnitude and correct for the restriction of range but they are also more difficult to interpret because each score stands for a different combination of item scores. Generally, the item-rest score correlations should reflect the progression of item difficulty expected by test developers. Children that can perform more developmentally advanced actions, reflected in items # 5 and #6 in each domain subscale, should also be able to complete the easier items.

The distribution of item responses confirms the test developer's expectation of low to medium item difficulty, as more than half of the milestone actions in gross-motor, fine-motor and personal-social domains can be completed by 95% of children. The communication domain has the highest proportion of difficult items, which the children at the concern categories (at-risk and monitoring) cannot complete, followed by the problem-solving domain. At the same time, only two domains, communication and gross motor, show the expected increase of difficulty with item rank number. In other domains, the items of medium and high difficulty are dispersed across the subscales suggesting that the developmental quotient is not an appropriate estimate of difficulty for this sample.

As for item discrimination, homogeneity of domain content seems to influence the magnitude of correlations along with item difficulty. For the item-rest score correlations, the inverse relationship between correlation magnitude and item difficulty holds for most items on the most difficult domains--

communication and problem solving. On other subscales, items of the same difficulty may have total score correlations ranging from 0.2 to 0.6 for the same item response pattern. While all correlation values indicate good discrimination (Nunnally and Bernstein (1994) suggest the range 0.15 to 0.3 as acceptable for such correlations), the exact link between magnitude and heterogeneity is not clear.

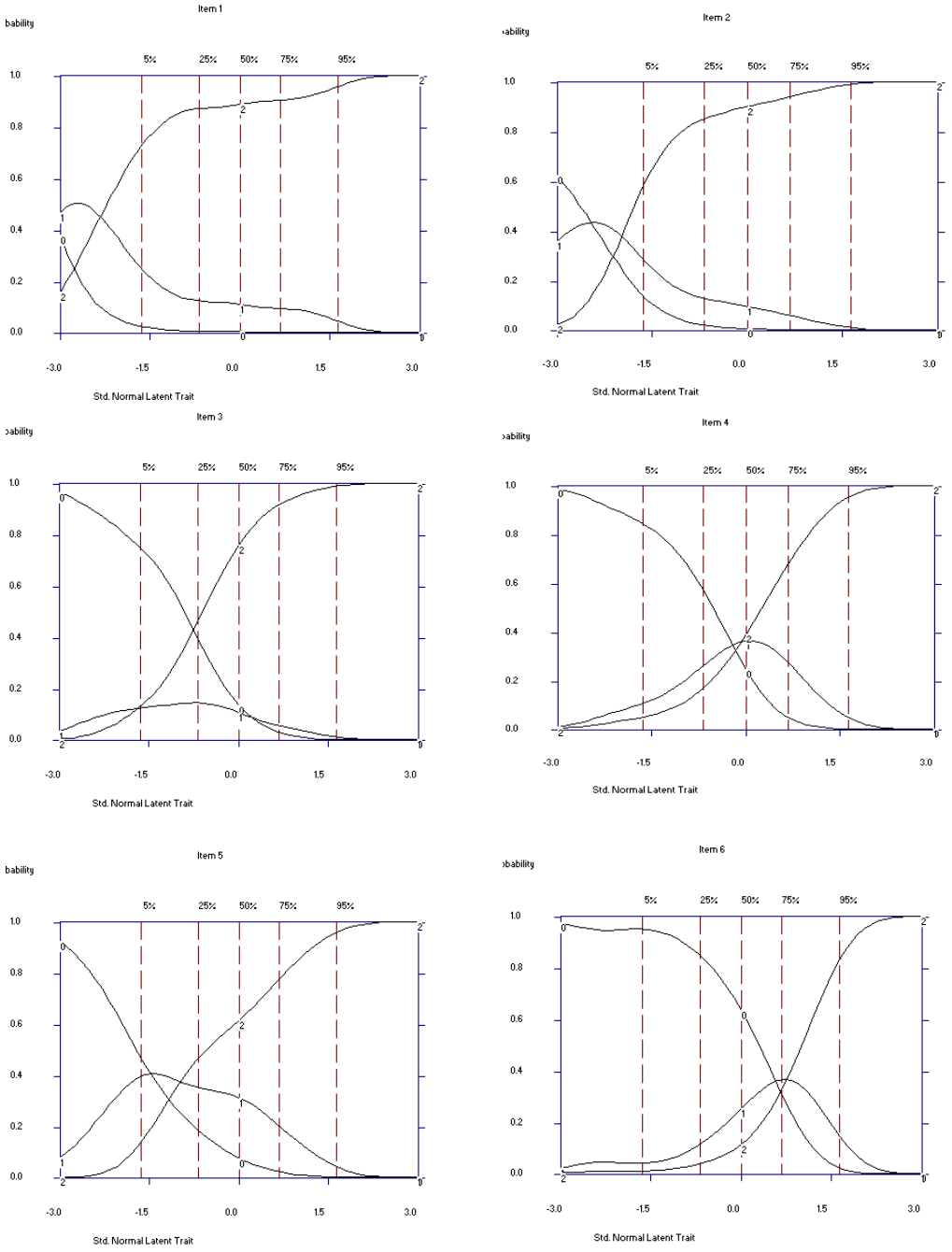
Data in Table 4 also allow for comparing observed score differences between children falling into each of the three categories, as well as difference in scores that could be observed with cut-off location matched with location in the normative sample distribution. In the Canadian observed score distribution of item responses, at-risk children differ from the monitoring zone children in that the former cannot fully complete all items on the motor domains, four items in the personal-social and problem-solving domains, and three items in the communication domain. Conversely, very few items show differences in responses between the children developing typically and children with some concerns (monitoring zone). Most domains have one to two items fulfilling those criteria, and only personal-social domain, with the monitoring zone cut-off located lower than expected, has four such items. In most cases, items are endorsed by children at the monitoring zone cut-off as well as by the children in the next percentile, with the exception of item #3 in the communication domain which children below the healthy development zone cannot endorse at all.

Relative difference in cut-off location is not associated with observed score difference for most separate items (22 out of 30). Such differences appear

for two items around the at-risk cut-off scores (C2) and for six items at the monitoring zone cut-off score (C1). In the case of gross-motor and personal-social domains, Canadian cut-off points are associated with lower scores enabling better separation of categories. The opposite can be seen in the problem-solving domain, where children at the 23rd percentile have the items fully endorsed, as do the children in the typical development zone.

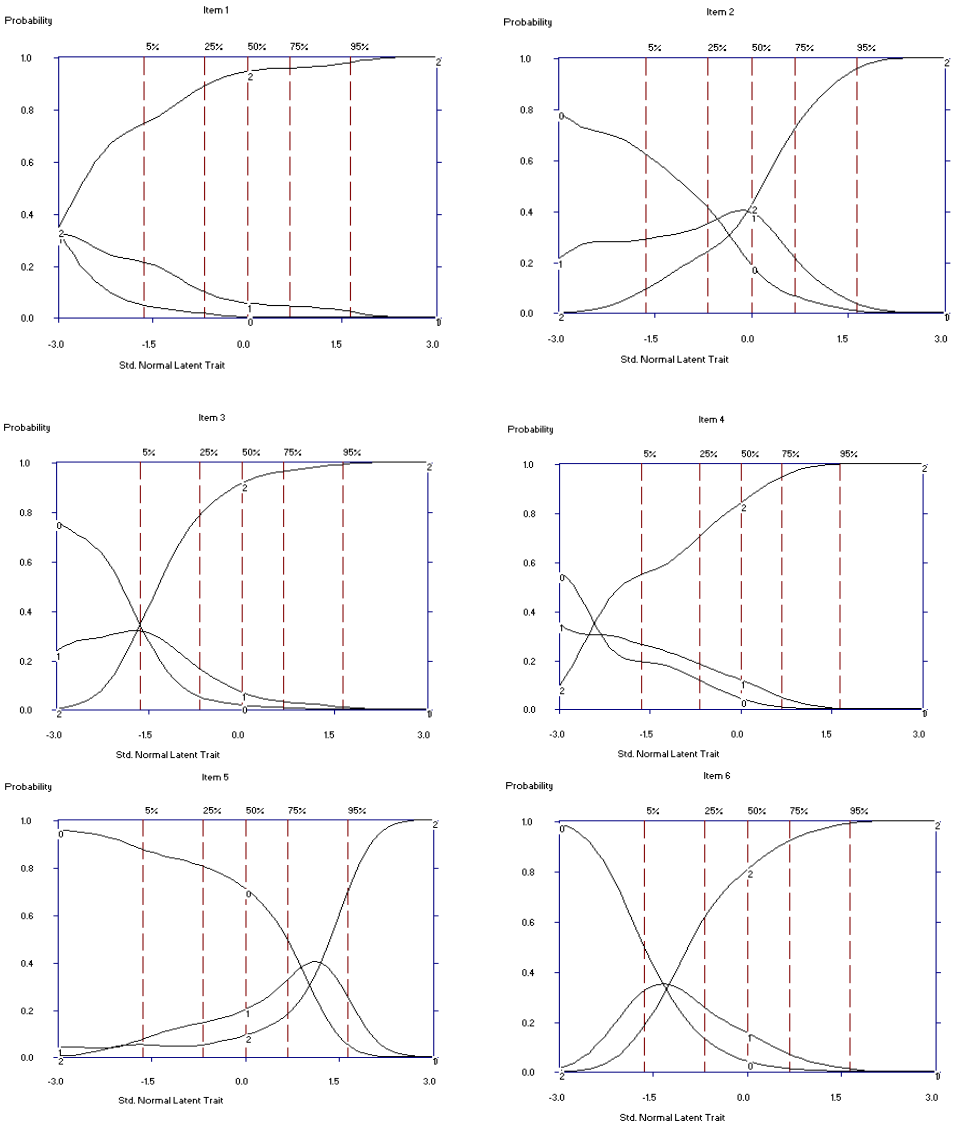
Option discrimination analysis. In the NIRT procedure, option characteristic curves model probabilities of selecting a certain option, conditional on the level of ability, and are basic building blocks for item and test-level discrimination and resulting information estimates. The option characteristic curves show how likely certain options are to be selected at specific ability levels and to what extent option weights reflect ability progression. The ability variable is reported in standard normal units to facilitate comparison across domains. Domain subscale descriptions are grouped based on difficulty, as items in more difficult and less difficult domains have similar discrimination patterns. More difficult domains, communication and problem-solving, are presented on Figures 4 and 5, respectively. Easier domains, fine-motor and personal-social, are presented on Figures 6 and 7, respectively.

Figure 4. Option characteristic curves for the communication domain



*Item options: ”0” – “Not Yet”, “1” – “Sometimes”, “2” – “Yes”

Figure 5. Option characteristic curves for the problem-solving domain



*Item options: ”0” – “Not Yet”, “1” – “Sometimes”, “2” – “Yes”

Difficulty and discrimination patterns for the communication and problem-solving items are similar and can be summarized as follows:

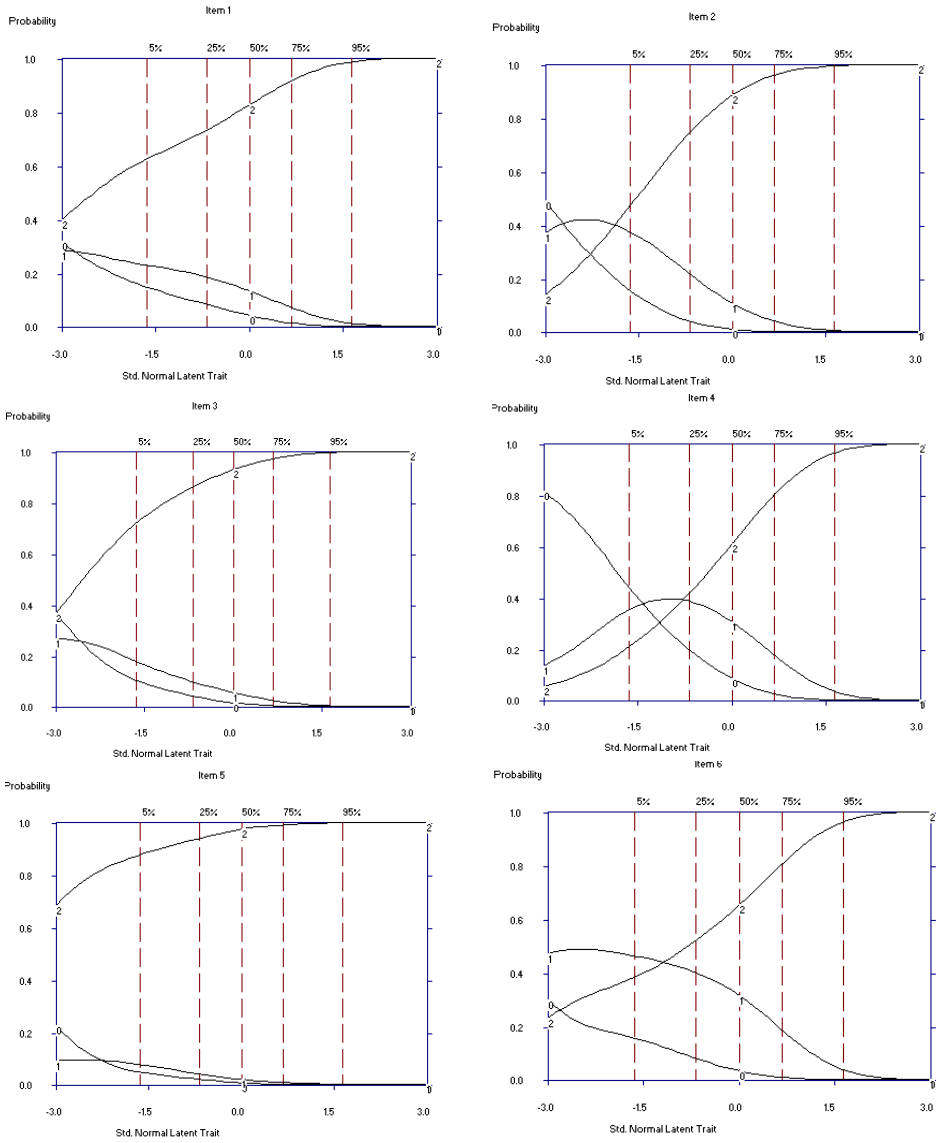
1. The easiest items, which are either fully endorsed (option 2/“yes”) for children at all ability levels or have all options crossing around the at-risk cut-off point

or lower. The communication domain has two such items (#2 and #3) and the problem-solving domain has three of them (#1, #3, #4)

2. Some items are very difficult for the target population, as they can only be endorsed by children at above-average levels of ability. Those are items #4 and #6 in the communication domain and item #5 in the problem-solving domain.
3. The remaining items in both domains are clearly effective in separating children with ability levels around at-risk cut-off from children with ability levels corresponding to the healthy development category. The former cannot endorse the items (who mostly have no items endorsed (option 0/"no"), while the latter have the items fully endorsed.
4. Children with ability levels falling into the monitoring zone (between the two cut-off scores) do not have a clear option combination pattern associated with their ability levels. For items other than those described in p.1-2, either each option has an equal opportunity of being selected at the cut-off between the monitoring and health development zone or children with these ability levels have the items fully endorsed (i.e., option 2 has the highest probability of being selected). For some items--#2 in the problem-solving domain, #4 on the communication domain--children with the ability levels in the monitoring zone and around the C1 cut-off scores are more likely to not have the item endorsed, i.e., their responses are similar to those children with the lowest ability levels.

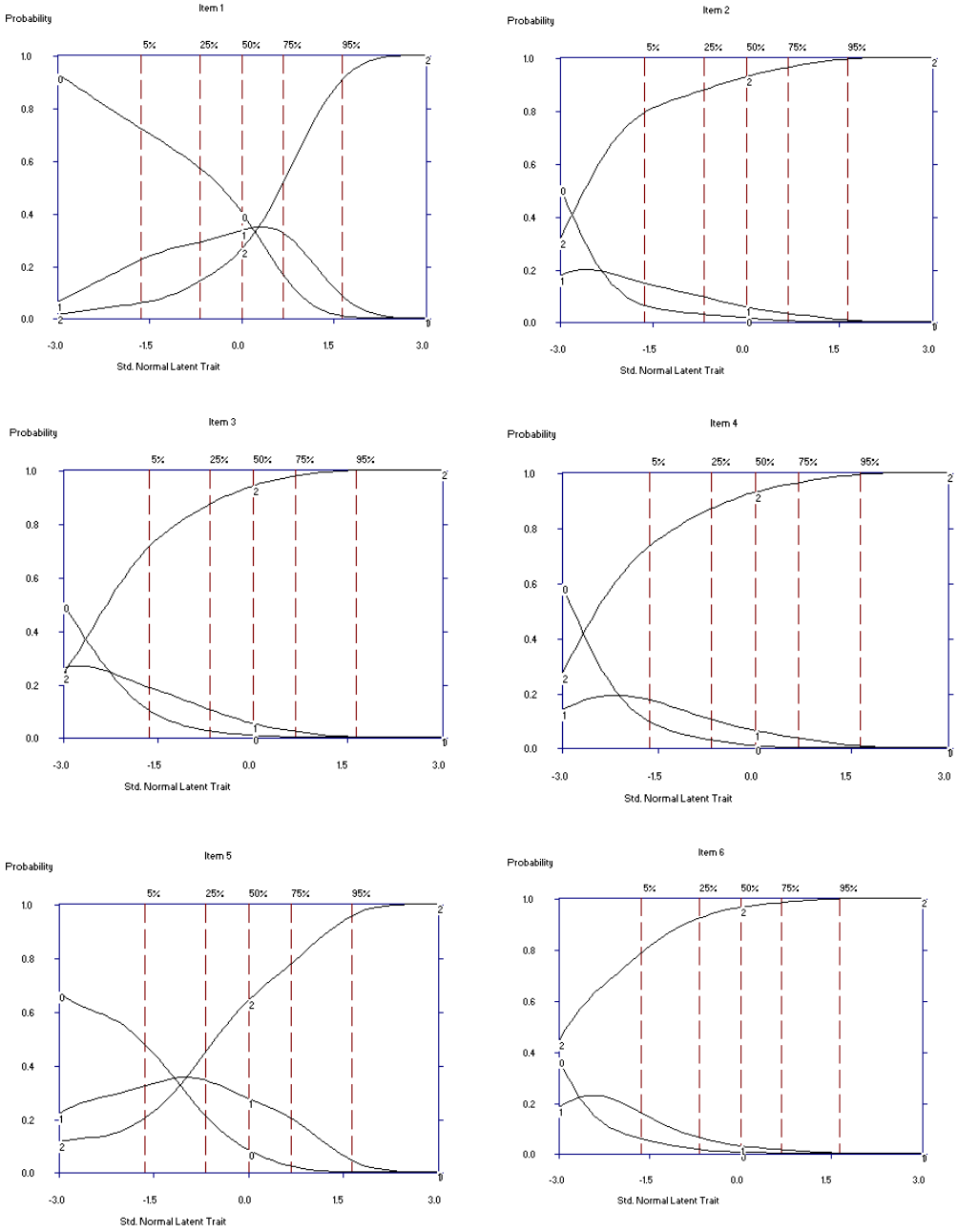
One of the reasons for a lack of item discrimination around the monitoring zone cut-off is limited discrimination power of option 1/"sometimes", which does not reflect an intermediate proficiency between "not yet" and "yes", as expected by test developers. It is rarely endorsed at any ability level or is mostly endorsed by children with lowest levels of ability, as is the case with item # 1 on the communication domain. As a result, monitoring zone children have higher probability of having low-medium items fully endorsed: i.e., they show the same response pattern as children with higher levels of ability. This outcome can potentially result in overidentification of some children on the problem-solving domain, as its monitoring zone category does not exceed 10 observed score points and one item score can make a difference in where the child will be placed. Lack of test information around the monitoring cut-off score for the problem-solving subscale seems to confirm the possibility of such a misclassification. The communication domain monitoring zone extends over 20 observed score points, which may help prevent undesired misclassifications.

Figure 6. Option characteristic curves for the fine motor domain



*Item options: ”0” – “Not Yet”, “1” – “Sometimes”, “2” – “Yes”

*Figure 7.*Option characteristic curves for the personal-social domain



*Item options: ”0” – “Not Yet”, “1” – “Sometimes”, “2” – “Yes”

As expected from the observed scores review, most items on these domains have a high probability of being fully endorsed by children at all ability levels. At least three items in each domain show options crossing around the at-risk cut-off score indicating equal probability of being selected and, thus, imprecision of the subscale around this cut-off. The remaining items differ in the following ways:

1. In the personal-social domain, #1 is very difficult for all children at the below-average ability level to complete while #5 is similar to any low-medium difficulty item in the other domains: not endorsed for at-risk ability levels and fully endorsed across the remaining ability scale. The high information around the cut-off score, shown by TIF analysis for the personal-social domain, contradicts the results of the option analysis. This discrepancy can either be explained by the inaccuracy of the parametric model due to the very low sample size (n=15) or by the strong contribution of the most difficult item across all subscales.
2. All fine motor domain items are relatively easy to complete for children in both concern categories. This is demonstrated by low probability of endorsing the option 0 ("Not Yet") at any ability level for 5 out of 6 domain items. Items # 2, # 4 and # 6 provide some discrimination around the at-risk cut-off (c2), where children at the lower levels of ability can "sometimes" (option 5) endorse an item, while stronger children definitely endorse them (option 10). In addition, all options cross around the monitoring zone cut-off (c1).

In sum, the analysis of option characteristic curves and their association with the ability level suggests that the pattern of progressive difficulty does not hold for most domains, as many items are too easy for children to complete and the "sometimes" option is rarely endorsed regardless of item difficulty. In addition, both observed item responses and option discrimination analysis suggest that most items work well for identifying children at the lowest levels of ability who can complete very few items. However, they are less effective in separating children in need for monitoring from those showing healthy development. Depending on the cut-off location on the subscale, potential for both underidentification (due to easiness of the items) and overidentification (due to low use of the heavily weighted "sometimes" option) exists for children at medium-low ability levels.

Chapter 5: Discussion and Conclusions

The purpose of the present study was to investigate the psychometric properties of ASQ-18 to provide information about tool's potential to identify children at-risk in the Canadian context. A combination of CTT and non-parametric IRT methods of item and subscale analysis were used to make conclusions about this potential using (a) analysis of reliability and error of measurement; (b) subscale information function and score precision; and (c) item difficulty and discrimination. The results at the domain level and item level are summarized below. They are followed by reviewing the relevance of findings for judging decision consistency of the scale, overview of limitations, and implications for research and practice.

Domain-level Evidence

Relative cut-off location. Despite the proximity of means and standard deviations in the two samples, location of the distribution-based cut-off scores in the Canadian sample differs for three out of five domains, leading to difference in observed scores falling immediately below the cut-offs. Cut-offs are located further from the mean in two domains and closer to the mean in one domain compared to the normative sample and these cut-off scores do not correspond to the percentiles of the normal distribution as they do in the US sample. Although the observed proportions for the US normative distribution are not known, it can be estimated, with the normal distribution percentiles as a reference that about 1% of children may be misclassified around the at-risk cut-off for most domains, with the exception of the problem-solving domain (estimated 7%). For children with

scores around the monitoring zone cut-off, the proportion of misclassified children is estimated to be around 7 to 10.

Reliability, preciseness and error of measurement. Cronbach's alpha values for each domain based on polychoric correlations were acceptable for all domains. However, the standard error of measurement resulting from this reliability estimate produced 95% confidence intervals that fully covered the monitoring zone in the case of at-risk cut-offs and partially covered the typical development zone in the case of the monitoring zone cut-offs. The conditional standard error of measurement for the observed cut-off scores, determined in part by a response pattern associated with each observed score, also spanned the monitoring zone and, for some response patterns, not only overlapped the monitoring zone between two cut-off scores but also included parts of the at-risk and typical development zones for some domains. The confidence intervals using the multinomial CSEM have not been calculated, as the procedure is still in development, but the findings from both methods indicate the possibility of misclassifications around both cut-off scores.

Another estimate of CSEM used in this study is the reciprocal of the NIRT information function of the maximum likelihood ability estimates. It shows that communication and problem-solving domains provide more information at the lower levels of the ability scale than do fine motor and personal social domain, although the latter domain has the maximum information around the at-risk cut-off. Fine motor domain provides less information at all levels of the ability scale than other three domains for which the TIF could be estimated. The 95%

confidence interval constructed using this type of CSEM confirms the results gained with other methods: intervals for the at-risk cut-off cover the monitoring zone while the CI for the monitoring zone cut-off span 35% to 50% of the ability scores in the typical development zone. For domains with less information and, consequently, less discrimination at the target ability levels, confidence intervals at each cut-off score span ability values falling into each of the three categories available, thus indicating a 95% probability of misclassification in either direction.

Item-level evidence

Several factors contributed to item difficulty and discrimination which, in turn, underlie precision of the ability estimates: expected age for the milestones, complexity of tasks and ability differences within the sample. The age-appropriateness of the items is expressed in "developmental quotients"; it has been verified against milestone charts and research literature findings, along with task complexity. For example, problem-solving and personal-social domain items all tap into several dimensions of cognitive and socio-emotional development and require integration of skills from different areas. As for ability variation, differences between Canadian general primary care population and the US normative sample, partially recruited through early intervention program were to be expected.

Item-level results confirmed the low to medium level of difficulty of most items: 3 to 4 items in each domain cover the milestones that should be achieved by the age of 18 months, even given the intra-individual variation. This finding

supports the purpose of the scale – to identify children not reaching important milestones by the maximum expected age. Empirical item difficulty did not match the expected order of difficulty measured by developmental quotients. However, it aligned with the description of the developmental milestones found in the literature. That is, tasks expected to be performed by 18 months of age were more often fully endorsed than tasks expected from the majority of children between 18 and 24 months of age.

Item discrimination provides further evidence that at least half of the items on each subscale were fully endorsed by children at the below-average ability levels. Option discrimination analysis in four domains has shown that in two domains with more difficult items – communication and problem solving – most items provide sufficient discrimination around the at-risk cut-off. In the easier domains – fine-motor and personal social – 4-5 out of 6 items are likely to be fully endorsed at all levels of ability, reducing the precision of ability estimates at the domain level. There seems to be no direct link between empirical item quality and conceptual complexity of tasks in terms of skills integration which is reflected in different amounts of information around the cut-offs on two complex domains – personal-social and problem-solving.

Implications for decision consistency with the US-set cut-off scores

These results highlight several factors that impact the consistency of classification in the new sample, although the exact extent and direction of misclassification is not known. Out of all the domains, communication should best separate at-risk from typically developing ones with the most consistency, as

it contains (a) a mix of more and less difficult items; (b) appropriate discrimination around the at-risk cut-off resulting in highest possible information and narrower 95% CI for score precision; (c) cut-off location that matches that of the normative sample; (d) a wide interval of 15 observed scores between the two cut-off scores, accounting for the large confidence interval around the at-risk cut-off. By comparison, scores in the personal-social domain are the least likely to be consistent around the at-risk cut-off because of low difficulty and discrimination of most items resulting in lower score precision along with a smaller interval (10 points) between the two cut-offs. In this domain, the observed cut-off scores also fall further from the mean and below the point of minus 2 and 1 standard deviations. This outcome can lead to underidentification of children with scores around both cut-offs.

For other domains, item and domain level evidence is somewhat contradictory and therefore more challenging to interpret. For the gross-motor domain, observed statistics suggest that underidentification of some children with lower levels of ability is possible, despite the fact that item difficulty and complexity of tasks matched the expectation of test developers and unidimensionality of this domain resulted in higher internal consistency estimates. Items in the problem-solving domain also showed appropriate preciseness around the at-risk cut-off due to better-discriminating items and higher amount of information at the at-risk levels of ability. However, difference in distributional location due to different standard deviations can lead to overidentification of children with some scores falling into the monitoring and typical development

zones for the problem-solving domain. For the fine motor domain, cut-off locations are identical in two distributions, but the low item difficulty and discrimination decreases the amount of information available even at the lowest levels of ability and increases the probability of a child with certain scores being classified in each of the three categories available.

Decision consistency in the "monitoring zone" between the at-risk and monitoring/typical development cut-offs raises questions about the status of that category. Regardless of score reliability and cut-off location, statistical models applied in this study suggest a 95% chance that any observed score below 1 standard deviation from the mean has a corresponding true ability score falling into either the at-risk group or monitoring zone. Similarly, some observed scores in the typical development zone can have a corresponding true score falling into the ability range that requires monitoring. The test manual does not specify clearly if the monitoring zone is a special category on the continuum, with certain risks associated with the scores, or if scores falling into this zone should be interpreted as scores likely indicating typical development. As a result, follow-up studies are still required to address this question. Some researchers have previously called for a specific monitoring category on the developmental screening tests, as children marginally passing developmental screening tests still display functioning problems in one or more domain (e.g., Glascoe, 2001; Pool & Hourcade, 2011). Results from this study confirm that in a general Canadian population, the ASQ-18 can be less effective in separating typically developing children from those who need follow-up and support.

This study also highlights two related problems that can influence consistency and accuracy of screening results in the new context: mode of administration and the scoring rule. It has been observed in a clinic setting that parents tend to complete the questionnaire from memory, without trying tasks out with their children (Rydz et al., 2006; Marks, 2007; Hix-Small, Marks, Squires, & Nickel, 2007). Although no detailed information was available on how exactly parents completed the questionnaire that provided data for this study, the results of item analysis seem to support that completion for memory was the case, even if it goes against the procedures recommended by the test manual (Squires et al., 2009). First, most items in each domain function as two-option items (yes-not yet), and the option "somewhat" has a very low probability of being selected for most items. Second, the items expected to be completed with and without demonstration (i.e. item # 4 and # 6 in the problem-solving domain) or items measuring conceptually similar tasks (items # 2-4 in the personal-social domain) show identical option response patterns, although these items were expected to display various degrees of difficulty. Completion of items from memory could potentially lead to inflated ratings for some items, especially for the complex tasks that rely on imitation and copying.

The potential threat to validity of classification, coming from the mode of questionnaire completion, can possibly be exacerbated by the scoring rule: large interval between response options (0, 5, and 10) and unweighted summation of item scores into subscale score. Use of the pre-determined cut-off scores, without regard to the sample standard deviation, resulted in a 10-point distance between

two cut-offs in 4 out of 5 domains, which would have been wider had the cut-offs been adjusted to the sample distribution. The narrow interval means that two items erroneously scored as "yes" instead of "sometimes" would place a child into the typical development zone instead of monitoring zone. This type of scoring problem could produce a systematic measurement error in addition to random error that is expected to be high for short subscales (Emons et.al., 2007) and might have contributed to lower precision estimates around the monitoring zone cut-off in the current study.

Limitations

Data source and model restrictions are the main limitations that influenced the design of the study and the extent of information about decision consistency and accuracy.

The data were collected for the purposes of service monitoring and evaluation, not psychometric validation. As a result some crucial variables (e.g., criterion variables clearly linked to each domain, environmental and psychosocial risk factors) were either not present or not sufficiently explained to be usable. In particular, the absence of the criterion precluded any definitive conclusions about probability and direction of misclassification for the existing cut-off scores. Most sample characteristics, item-level results and some domain-level statistics were not available in the original test manual. This omission makes comparisons for these variables impossible.

While internal consistency was the only possible CTT reliability measure that could be estimated for the data, items may not the most important source of

measurement error in the screening context. In addition, internal consistency will be lower in some domains with complex tasks requiring integration of skills from various domains. The CSEM estimates for observed scores corresponding to cut-offs complemented the average SEM from CTT reliability coefficients, but the multinomial model underlying the CSEM has its own limitations. First, the multinomial curves are known to approach 0 at the extremes, although this does not mean absence of error variation. Second, the response patterns, and not the observed scores per se, are sources of error, leading to a variety of possible CSEM values for each observed score. Third, the CSEM is expected to be larger, than CTT SEM, both methods are not directly comparable for the same observed scores. In addition, this model has been rarely used with real data (Lee, 2005).

NIRT based on Gaussian kernel smoothing also has a number of important limitations. First, due to the local averaging technique, the curves are difficult to summarize in a single difficulty and discrimination value, as it is the case with the parametric models. Second, the parametric estimation of the test information function differs from the approach used in item analysis and the accuracy of the logistic-quadratic model has not been extensively tested (Ramsay, 2001). Third, the value of the bandwidth parameter, used in local averaging, is crucial for determining how well the smoothed curve will approximate the data. In the case of study samples, the recommended bandwidth parameter value had to be increased 1.5 to 2 times to overcome monotonicity problems and the preliminary visual inspection of the graphs revealed considerable deviations from the real data shape for some very easy items. Fourth, Gaussian kernel is the only kernel

function offered by the Testgraf software. While we assume underlying normal distribution of developmental ability, the observed scores are negatively skewed.

Implications for research and practice

Implications for research. The findings of this exploratory study regarding potential problems with decision consistency and accuracy should be verified with a full diagnostic accuracy study, which should include each of the five domains with regards to sensitivity, specificity and positive predictive value of the classification results against specific standardized diagnostic assessment. Examining mode of administration (e.g., completion in the clinic versus at home, with or without help, from memory versus trying tasks out) and its effect on rating accuracy is also recommended. Item – and domain level psychometric properties relevant for consistency and accuracy of classifications can also be verified by finding a fitting parametric IRT model which could produce population-independent parameters and which would align with the newer computer-based version of the ASQ-18. Decision regarding development of Canadian or local norms, or adjustment of scoring rule, can be made based on the results of these studies.

Implications for practice. Brief general developmental screeners are reported to be most effective when used at multiple time points and combined with routine pediatric surveillance and other ways to elicit parents' concerns (American Academy of Pediatrics, 2006; Marks & La Rosa, 2012). The results from this study support these recommendations in that multiple sources of random and systematic error, combined with large number of developmental change in

toddlers, can produce inaccurate results at a single time point and potentially missing some delays. It is also recommended that the DQ values, adapted from other tests, be verified against the recent research evidence on child development (Bremner & Wachs, 2010). It may also be necessary to adjust the observed scores, corresponding to the cut-off scores for three domains by taking the next observed score above or below the recommended cut-off.

References

Accardo, P. J. & Whitman, B. Y. (2005). *Dictionary of developmental disabilities terminology*. Baltimore, Md.: Paul H. Brookes.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: AERA, APA & NCME.

Aylward, G. P. (1997). Conceptual issues in developmental screening and assessment. *Journal of Developmental & Behavioral Pediatrics*, *18*(5), 340–349.

Aylward, G. P. (2005). The conundrum of prediction. *Pediatrics*, *116*(2), 491–492.

Aylward, G. P. (2009). Developmental screening and assessment: what are we thinking? *Journal of Developmental & Behavioral Pediatrics*, *30*(2), 169–173.

Aylward, G. & Stancin, T. (2008). Developmental screening and assessment: measurement and psychometric considerations. In M. L. Wolraich, D.D. Drotar, P.H. Dworkin, & E.C. Perrin (Eds.), *Developmental-behavioral pediatrics: evidence and practice* (pp. 123–130). Philadelphia: Elsevier.

Bellman, M., Byrne, O. & Sege, R. (2013). Developmental assessment of children. *BMJ*, *15(346)*. doi:10.1136/bmj.e8687

Berger, S. P., Hopkins, J., Bae, H., Hella, B. & Strickland, J. (2010). Infant Assessment. In G.J. Bremner & T.D. Wachs (Eds), *The Wiley-Blackwell handbook of child development* (pp. 226–256). Malden, MA: Wiley-Blackwell.

Camilli, G. (2006). Test fairness. In R. Brennan (Ed.), *Educational measurement* (pp. 221–256). Westport, CT: Praeger .

Camp, B. W. (2006). What the clinician really needs to know: questioning the clinical usefulness of sensitivity and specificity in studies of screening tests. *Journal of Developmental & Behavioral Pediatrics*, *27*(3), 226–230.

Camp, B. W. (2007). Evaluating bias in validity studies of developmental/behavioral screening tests. *Journal of Developmental & Behavioral Pediatrics*, *28*(3), 234–240.

Christ, T.A., and Nelson, P.M. (2014). Developing and evaluating screening systems: practical and psychometric considerations. In: R.K. Kettler, T.A. Glover, C.A. Albers, & K. Feeney-Kettler (Eds.), *Universal screening in*

educational settings. Evidence-based decision-making for schools. Washington, DC: American Psychological Association.

Drotar, D., Stancin, T., Dworkin, P. H., Sices, L. & Wood, S. (2008). Selecting developmental surveillance and screening tools. *Pediatrics in Review*, *29*(10), 52–58.

Eliot, L. (1999). *What is going on in there? How the brain and mind develop in the first five years of life.* New York, NY: Bantam.

Embretson, S. E. & Reise, P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Emons, W. H., Sijtsma, K. & Meijer, R. R. (2007). On the consistency of individual classification using short scales. *Psychological Methods*, *12*(1), 105-120.

Feldman, H. & Messick, C. (2007). Assessment of speech and language. In M. L. Wolraich, D.D. Drotar, P.H. Dworkin, & E.C. Perrin (Eds.), *Developmental-behavioral pediatrics: evidence and practice* (pp. 177–190). Philadelphia: Elsevier.

Frisk, V., Montgomery, L., Boychyn, E., Young, R., McLachlan, D., Neufeld, J. & others. (2009). Why screening Canadian preschoolers for language delays is more difficult than it should be. *Infants & Young Children*, *22*(4), 290–308.

Gadermann, A. M., Guhn, M. & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research & Evaluation*, *17*(3), 1–13.

Glascoe, Frances Page. (2001). Are overreferrals on developmental screening tests really a problem? *Archives of Pediatrics & Adolescent Medicine*, *155*(1), 54–59.

Glascoe, Frances Page. (2005). Screening for developmental and behavioral problems. *Mental Retardation and Developmental Disabilities Research Reviews*, *11*(3), 173–179.

Glascoe, Frances P, & Dworkin, P. H. (2008). Surveillance and Screening for Development and Behavior. In M. L. Wolraich, D.D. Drotar, P.H. Dworkin, & E.C. Perrin (Eds.), *Developmental-behavioral pediatrics: evidence and practice* (pp. 130–144). Philadelphia: Elsevier.

Gokiert, R. J., Chow, W., Parsa, B., Rajani, N., Bisanz, J. & Vanderberghe, C. (2010). Early childhood screening in immigrant and refugee populations.

Edmonton, Alberta, Canada: Community-University Partnership for the Study of Children, Youth, and Families.

Gokiert, R. J., Georgis, R., Tremblay, M., Krishnan, V., Vandenberghe, C., Lee, C. (2014). Evaluating the Adequacy of Social-Emotional Measures in Early Childhood. *Journal of Psychoeducational Assessment*, doi: 0734282913516718.

Hetherington, E. M., Parke, R. & Locke, V. O. (2006). *Child Psychology: A Contemporary Viewpoint*. Boston, MA: McGraw-Hill.

Hix-Small, H., Marks, K., Squires, J. & Nickel, R. (2007). Impact of implementing developmental screening at 12 and 24 months in a pediatric practice. *Pediatrics*, *120*(2), 381–389.

Jöreskog, K. G. & Sörbom, D. (1996). *PRELIS 2 User's Reference Guide: A Program for Multivariate Data Screening and Data Summarization: a Preprocessor for LISREL*. Scientific Software International.

Kerstjens, J. M., Bos, A. F., ten Vergert, E. M., de Meer, G., Butcher, P. R. & Reijneveld, S. A. (2009). Support for the global feasibility of the Ages and Stages Questionnaire as developmental screener. *Early Human Development*, *85*(7), 443–447.

Lee, W. (2005). A multinomial error model for tests with polytomous items (CASMA Research Report No. 10). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment.

Limbos, M. M. & Joyce, D. P. (2011). Comparison of the ASQ and PEDS in screening for developmental delay in children presenting for primary care. *Journal of Developmental & Behavioral Pediatrics*, *32*(7), 499–511.

Livingston, S. A. (2006). Item analysis. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of test development*, 421–441. Mahwah, NJ: Erlbaum

Macy, M. (2012). The evidence behind developmental screening instruments. *Infants & Young Children*, *25*(1), 19–61.

Marks, K. (2007). Should general paediatricians not select the ASQ in light of the Rydz et. al. study? *Pediatrics,120;457-458.*

Marks, K. P. & LaRosa, A. C. (2012). Understanding developmental-behavioral screening measures. *Pediatrics in Review*, *33*(10), 448–458.

Metzloff, A. N. & Williamson, R. A. (2010). The importance of imitation for theories of social-cognitive developent. In G.J. Bremner & T.D. Wachs (Eds),

The Wiley-Blackwell handbook of child development (pp. 345–364). Malden, MA: Wiley-Blackwell.

Nunnally, J. C. & Bernstein, I. H. (1994). *Psychometric theory.* New York: McGraw-Hill.

Papalia, D., Olds, S., Feldman, R. & Kruk, R. (2008). *A Child's World: Infancy through Adolescence*. New York: McGraw-Hill Ryerson.

Pool, J. L. & Hourcade, J. J. (2011). Developmental screening: A review of contemporary practice. *Education and Training in Autism and Developmental Disabilities*, *46*(2), 267.

Ramsay, J. (2001). TESTGRAF manual. *Montreal, Quebec, Canada: McGill University*.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, *56*(4), 611–630.

Reise, S. P. & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*, 27–48.

Rochat, P. (2010). Emerging Self-Concept. In G.J. Bremner & T.D. Wachs (Eds), *The Wiley-Blackwell handbook of child development* (pp. 320–345). Malden, MA: Wiley-Blackwell.

Ross, H., Vickar, M. & Perlman, M. (2010). Early social cognitive skills at play in toddlers' peer interaction. In G.J. Bremner & T.D. Wachs (Eds), *The Wiley-Blackwell handbook of child development* (pp. 510–532). Malden, MA: Wiley-Blackwell.

Salvia, J. & Ysseldyke, J. (2009). *Assessment in special and inclusive education*. Cengage Learning.

Santor, D. A. (2005). Using and evaluating psychometric measures.In J. Miles &P. Gilbert (Eds.), *Handbook of research methods for clinical and health psychology* (pp. 95–101).Oxford: Oxford University Press.

Santor, D. A., Ascher-Svanum, H., Lindenmayer, J.-P. & Obenchain, R. L. (2007). Item response analysis of the Positive and Negative Syndrome Scale. *BMC psychiatry*, *7*(1), 66-71

Santor, D. A. & Coyne, J. C. (2001). Examining symptom expression as a function of symptom severity: Item performance on the Hamilton Rating Scale for Depression. *Psychological Assessment*, *13*(1), 127-139.

Schonhaut, L., Armijo, I., Schonstedt, M., Alvarez, J. & Cordero, M. (2013). Validity of the ages and stages questionnaires in term and preterm infants. *Pediatrics*, *131*(5), 1468–74.

Simard, M.-N., Luu, T. M. & Gosselin, J. (2012). Concurrent validity of ages and stages questionnaires in preterm infants. *Pediatrics*, *130*(1), 108–114.

Sonnander, K. (2000). Early identification of children with developmental disabilities. *Acta Paediatrica*, *89*(434), 17–23.

Squires, J., Twombly, E., Bricker, D. D. & Potter, L. (2009). *Ages & Stages Questionnaires: A Parent-Completed Child-Monitoring System*. Baltimore, Md.: Paul H. Brookes.

Streiner, D. L. & Norman, G. R. (2008). *Health measurement scales: a practical guide to their development and use*. Oxford: Oxford University Press.

Yen, W. M. & Fitzpatrick, A. R. (2006). Item response theory. In R. Brennan (Ed.), *Educational measurement* (pp. 111–153). Westport, CT: Praeger .

Yovanoff, P., Squires, J. & McManus, S. (2013). Adaptation from pencil-paper to web-based administration of a developmental questionnaire for young children. *Infants & Young Children*, *26*(4), 318-332.

Yovanoff, P., & Squires, J. (2006). Determining cutoff scores on a developmental screening measure: Use of receiver operating characteristics and item response theory. *Journal of Early Intervention*, *29*(1), 48–62.

Zumbo, B. D., Gadermann, A. M. & Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods*, *6*(1), 21–29.