A systematic review of measurement instruments to assess cognition and language development at 24 months of age, for use in effectiveness trials of nurse-home visitation programs

Jennifer Lam, BHK, Master of Public Health Candidate

Faculty of Health Sciences

Simon Fraser University

Senior Supervisor & First Reader: Dr. Nicole Berry

Second Reader: Dr. Nicole Catherine

July 30, 2015

Abstract

This systematic review evaluates cognitive and language measurement instruments for use at 24 months of age in effectiveness trials of nurse-home visitation programs. In particular, this review aims to identify and recommend potential instruments for the British Columbia Healthy Connections Project, a scientific evaluation of the Nurse Family Partnership, a nurse-home visitation program, in Canada. Although there is an overlap in child cognitive and language development in young children, the extent of the overlap is unclear, and hence it is recommended that instruments designed to separately assess cognition and language be used if feasible. A general search of potential instruments was completed, in addition to searches pertaining to instruments that have been used in home visitation interventions designed to improve language and cognition in young children.

Detailed components are reported for 6 instruments: the Bayley Scales of Infant and Toddler Development – Third Edition (Bayley-III), the Battelle Developmental Inventory – Second Edition (BDI-2), the Preschool Language Scale – Fifth Edition (PLS-5), the MacArthur-Bates Communicative Development Inventory (CDI), the Language Development Survey (LDS), and the Language Use Inventory for Young Children (LUI). All 6 instruments were considered as acceptable for reliability and validity ($r > 0.70$). Although the Bayley-III is considered the gold standard, without adequate resources and planning, it presents challenges in training and administration. The BDI-2 is a suitable substitute for the Bayley-III in lower resource situations. More research is required to draw conclusions on the reliability and validity of the PLS-5. Selection between the CDI, LDS, and LUI depends on what aspect of language development is to be evaluated. Child cognitive and language instruments administered at 24 months of age have limitations in their predictive validity and use in populations speaking English as a second language. Further research and longitudinal studies in these areas are warranted.

A systematic review of measurement instruments to assess child cognition and language development at 24 months of age, for use in nurse-home visitation programs

Early childhood cognitive and language development can be predictive of language and cognitive competence later in life (Ortiz-Mantilla, Choudhury, Leevers, & Benasich, 2008). Development has been observed to be impacted by adult-child interactions, where higher quality care during infant and toddler years can lead to better cognitive and language functioning in later years (National Institute of Child Health and Human Development Early Child Care Research Network, 2000). In addition, socioeconomic status (SES), social disadvantage, and other environmental factors in infant and toddler care can also greatly influence child cognitive and language development (Johnson & Marlow, 2006; Law & Roy, 2008). A vast number of interventions have been developed to assist families in overcoming challenges stemming from these factors, with notable effectiveness found in home visitation programs. Given the ability to provide services tailored for each family and home environment, home visitation programs have been shown to be beneficial to disadvantaged families for over 20 years for a range of outcomes, including child development (Peacock et al., 2013).

Two important areas of child development for which home visits have been studied are cognitive and language development. Cognitive development consists of attaining problem solving skills and using intuition, reasoning, and perception to learn new information and apply it to future situations (Rydz, Shevell, Majnemer, & Oskoui, 2005). Language development in early childhood consists of receptive (i.e., auditory, visual) and expressive (i.e., oral, verbal) language skills, in addition to articulation and nonverbal symbols (see Appendix D for detailed definitions). Receptive and expressive language development can be measured through a child's use of syntax, grammar, semantics, and vocabulary. Language development can also be assessed

through pragmatics. Pragmatics refers to the social use of language and focuses on language as a means to communicate with others, including skills such as adapting language for different purposes (Pesco & O'Neill, 2011).

There has long been an interest in measuring child cognitive and language development to further understand human developmental trajectories, as well as to identify developmental delays and disabilities for appropriate intervention (Deniz Can, Ginsburg-Block, Golinkoff, & Hirsh-Pasek, 2012). At 24 months of age, children are simultaneously developing cognitive and language skills, with evidence suggesting that cognitive development leads into the acquisition of language (MacNamara, 1972). Research suggests that children first develop non-linguistic cognitive processes before acquiring processes to use language. However, although cognitive and language development are correlated during childhood, there does not appear to be evidence for using the measurement of language development as a proxy to assess cognitive development, or vice versa (Siegel, 1981). This is likely due to an inability to quantify the extent of overlap between cognitive and language development, as well as the variability across children's developmental trajectories. As a result, a wide variety of psychometric child instruments have been developed, each using different testing methods (e.g., differences in testing environments, type of examiner and development domain tested, and assessment procedures) to specifically evaluate child cognitive or language development.

Child development instruments can be broadly categorized into two groups based on the method of data collection: 1) using external examiners' assessments, or 2) using parent-reports. The former, using standardized instruments administered by external examiners, provide norm-referenced scores based on data from the general population (Rescorla & Achenbach, 2002). While research has found comparable validity and normed scores in instruments relying on

parent-reports, parents represent a possible source of bias that the use of external examiners tend to eliminate (Johnson & Marlow, 2006; Law & Roy, 2008). However, given the short period of time an external examiner spends with a child, instruments using parent-reports have become increasingly common, as parents are able to report on regular, natural behaviour that may not be present during artificial testing periods (Law & Roy, 2008; Skarakis-Doyle, Campbell, & Dempsey, 2009). Instruments using parent-reports on child cognitive and language development have also been shown to be valid and reliable for child assessment in resource-limited environments (Rescorla & Achenbach, 2002; Johnson & Marlow, 2006; Deniz Can et al., 2012). As a result, many child development instruments contain at least a portion of parent reports (O'Neill, 2007).

Both parent-reported and examiner-administered instruments have been used as indicators of program effectiveness in child health programs and initiatives, including home visitation programs (Peacock et al., 2013). However, due to differences in areas including budgeting, available resources, testing environments, population demographics, and outcomes of interest, there is no single measure used to assess child cognition and language in home visitation programs. Hence, it is of importance to review the relevant literature to determine the most appropriate measure for examining the effectiveness of each home visitation program on child development.

The British Columbia Healthy Connections Project (BCHCP)

The BCHCP is a province-wide randomized controlled trial (RCT) evaluating the Nurse-Family Partnership program (NFP) in British Columbia (BC), Canada. NFP involves public health nurses providing intensive supports in the home to young women experiencing socioeconomic disadvantage who are pregnant for the first time – starting during the prenatal

period, and continuing until their children are two years old (Olds et al., 1986, 1997, 2002). NFP has been shown to improve mental and physical health outcomes for children and mothers in three United States (US) randomized controlled trials, but the program has never been rigorously evaluated in Canada. The RCT component of the BCHCP is therefore evaluating NFP's effectiveness compared to existing services on a variety of outcomes, including child cognitive and language development at 24 months of age. These outcomes will be assessed by BCHCP scientific field interviewers during in-person interviews with mothers and children at 24 months of age.

Purpose of the review

The current review has several goals, with a central purpose to evaluate the reliability and validity of various child cognition and language measurement instruments selected for consideration for the BCHCP. Careful consideration will be given to identify instruments that will be reliable and valid for the BCHCP study population – children born to young, first-time mothers experiencing socioeconomic disadvantage who are able to converse in English, but who may speak English as a second language (ESL). To contribute towards the BCHCP's defined outcomes as well as long-term follow-up goals, shortlisted instruments are assessed regarding their predictive validity and potential for use beyond 24 months of age. In addition to effectively measuring the constructs of interest, it is important to evaluate whether the scores from these instruments can be predictive of children's future cognition and language skills, particularly to see if NFP can have effects on school readiness. Staffing, administration, and budgetary considerations are also detailed for these instruments to determine their feasibility for the BCHCP. Although examiner-administered child assessment instruments are more likely to be objective, they are also more resource intensive. Given the budgetary considerations of any large

trial such as the BCHCP, parent-reported measures of child language development will also be a focus of the review. After taking into account the instruments used in previous NFP trials and other home visitation programs, as well as the validity and reliability of various instruments, this review reports on the top 3 examiner-administered child assessment instruments, and the top 3 parent-reported language instruments for use at 24 months of age.

Methods

A series of electronic searches of published articles was conducted using PsycINFO, PubMed, ScienceDirect, and Web of Science. The first phase of searches was a general search for early childhood cognition and language instruments. This was completed to identify a comprehensive list of potential measures. The focus of the general search was child development (related terms: preschool, infant), measurement (related terms: test, tool, assessment, scale, survey, inventory), cognition, language (related terms: speech, communication, linguistic), reliability, and validity (related terms: predictability). Appropriate search filters were applied to further specify the search topic. For example, the PubMed general search was completed as follows: ((((development* OR child* OR preschool OR infant)) AND (language* OR cognition*)) AND (speech OR communication* OR linguistic*)) AND (measure* OR test OR tool OR assessment* OR scale OR survey OR inventory) AND (validity OR reliability) Filters: Full text; published in the last 10 years; Humans; English; Child: birth-18 years; Infant: birth-23 months. Instruments used for other NFP trials were also searched. To find instruments used in comparable studies of home visitation programs and initiatives, the Home Visiting Evidence of Effectiveness (HomVEE: http://homvee.acf.hhs.gov) website was accessed. Articles were considered eligible for this review if the following inclusion criteria were met:

1. Article described at least one child cognition or language measurement instrument that has been validated for use in infants, toddlers, and preschool children.
2. Article described an instrument that is in English, or in English and other languages.
3. Article was published within January 2005 and May 2015.
4. Article described an instrument that could be used for the context of the BCHCP: families that are defined as disadvantaged, at-risk, or low income, or described an instrument that is designed for community use.

Articles describing instruments still in development, or instruments that are not commercially available were excluded. HomVEE-listed programs with missing data on child development measurement were also excluded from further review. It is of importance to note that the BCHCP uses a community sample (at risk for socio-economic disadvantage), and administers child assessments in the home environment of these families. Hence, articles describing instruments used for clinical populations (i.e. developmental disabilities, visual, motor, speech impairments, genetic diseases), or instruments designed only for use in laboratory or clinic settings were excluded. However, given the possibility that there may be children with clinical diagnoses in the BCHCP sample, the ability of the instruments to detect clinical risks were taken into account in the shortlisting of instruments. All potential instruments were also screened for the child development domains tested, eligible age group, data collection method, and administration details. Further, potential instruments from the general search were evaluated for their applicability for the BCHCP sample, and whether these instruments were able to find statistically significant differences in child development in other home visitation research studies (using HomVEE). For generalizability of the findings to existing NFP trials, potential instruments for the BCHCP were shortlisted if they were also used in other NFP RCTs.

The second phase of searches focused on the shortlisted instruments' reliability, validity, and use for populations that are defined as disadvantaged, at-risk, low income, or ESL. To capture as many articles on each shortlisted instrument, these searches were completed using the names of the instruments, as well as variations of their names (i.e. short-forms, acronyms). When appropriate, references of the resulting articles were explored. Searches in the grey literature were also completed to gather information on administration and budgetary considerations.

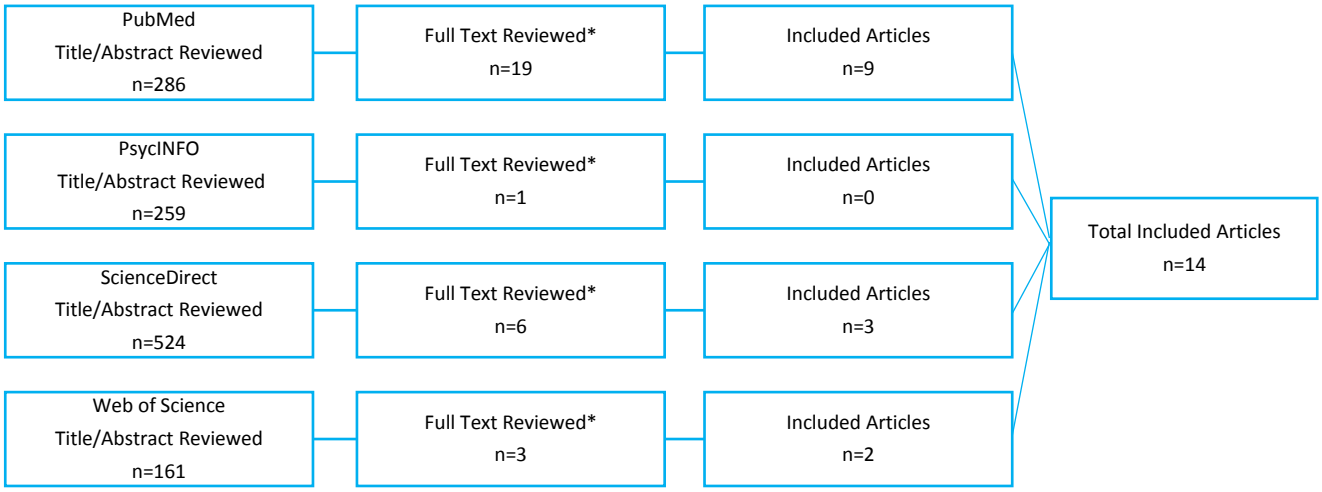
Results

Of the 1230 articles from the first general search (PubMed: 286, PsycINFO: 259, ScienceDirect: 524; Web of Science: 161), 14 articles were included in the analysis after screening for the inclusion criteria (Figure 1). These 14 articles reported on a wide range of different child development instruments for use in the general population. The complete list of potential instruments can be found in Appendix A. Both examiner-administered instruments and parent-reported instruments were shortlisted. The shortlisted instruments were chosen after assessing their potential use for the BCHCP context, and evaluating their use in eligible programs included in HomVEE (Appendix B), as well as other NFP trials (Appendix C). The search for further evidence for the six shortlisted instruments garnered a total of 25 articles (Figure 2). Some of these articles were duplicates from the general search, and a portion of these articles also reported on more than one instrument – hence the number of articles shown on Figure 2 may appear to be greater than 25.

Detailed descriptions of the six shortlisted instruments are outlined in Table 1. The shortlisted examiner-administered instruments were the Bayley Scales of Infant and Toddler Development – Third Edition (Bayley-III), the Battelle Developmental Inventory –Second Edition (BDI-2), and the Preschool Language Scale – Fifth Edition (PLS-5). The shortlisted

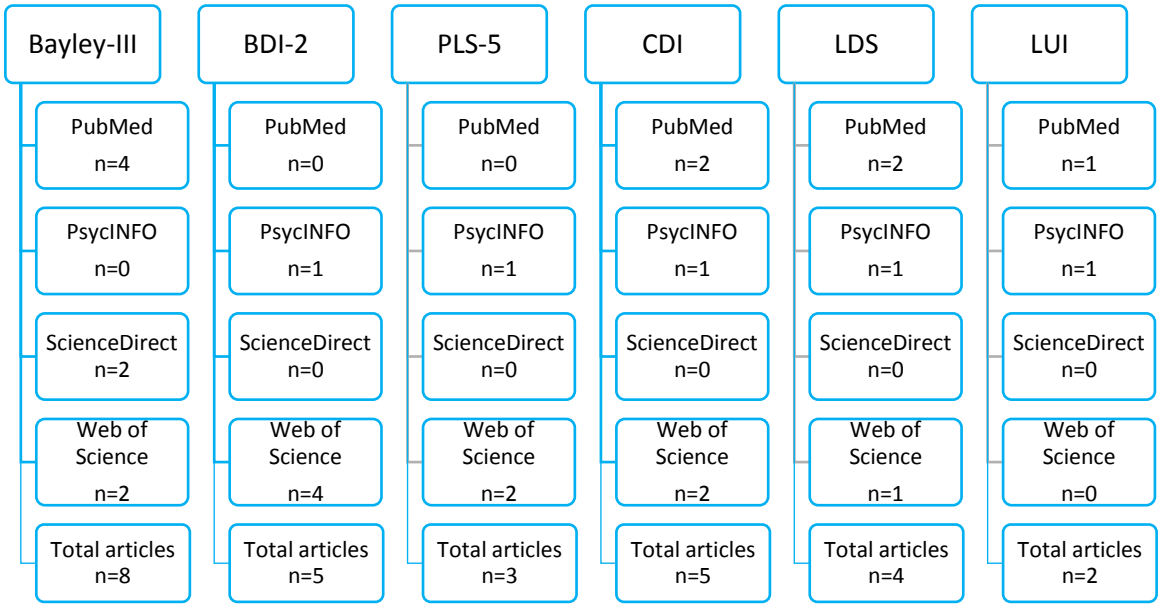
parent-reported instruments were the MacArthur-Bates Communicative Development Inventory (CDI), the Language Development Survey (LDS), and the Language Use Inventory for Young Children (LUI).

Figure 1. Flowchart of the systematic search and screening process.*



*After removing 19 duplicate articles found in the 4 databases

Figure 2. Flowchart of included articles for the shortlisted instruments.*



Bayley-III: Bayley Scales of Infant and Toddler Development, Third Edition; PLS-5: Preschool Language Scales, Fifth Edition; BDI-2: Battelle Developmental Inventory, Second Edition; LUI: Language Use Inventory for Young Children; CDI: MacArthur-Bates Communicative Development Inventories; LDS: Language Development Survey

*Reported numbers are totals after removing duplicate articles found from the 4 databases

Previous versions of the Bayley-III and PLS-5, as well as the current version of the CDI have been used to evaluate child cognitive and language development within research and evaluations of home visitation programs (Appendices B & C). The older versions of Bayley-III and PLS-5 were used in NFP trials in Denver, Memphis, and Elmira, as well as a number of home visitation programs, such as Healthy Families America and Parents as Teachers (Olds et al., 1986, 1997, 2002; Love et al., 2005; King et al., 2005). The CDI was also used in home visitation programs, including Early Head Start – Home Visiting and Healthy Steps (Love et al., 2001; Johnston et al., 2006). Although it was not included in the HomVEE database, the BDI-2 was used in Head Start, a pre-kindergarten program for low income families (Hallam et al., 2014). The current review did not yield results indicating the use of the LDS and LUI in evaluations of home visitation programs.

To evaluate the feasibility of these instruments for the BCHCP, thorough descriptions of the instruments are presented, in addition to administration and budgetary considerations for using these instruments for children at 24 months of age (Table 1, Appendix E). For instruments with capabilities beyond assessing cognitive and language development, specific administration and budgetary details are made to estimate the resources necessary to administer only the cognition and language portions, as these are outcomes of interest in the BCHCP. Each instrument's reliability to provide consistent results, as well as its validity to measure the intended domains are also be evaluated. Coefficients above 0.70 are considered acceptable, with higher values indicating a stronger instrument. Specific definitions pertaining to child development domains, as well as interpretations of reliability and validity coefficients can be found in Appendix D.

Examiner-administered instruments for cognition and language

Bayley Scales of Infant and Toddler Development, Third Edition (Bayley-III)

Description. The Bayley-III is regarded as the gold standard in child development assessment, and is often used as the instrument to validate other instruments (Lung et al., 2009). It assesses children 1-42 months of age in 5 domains: cognitive, language, motor, social-emotional, and adaptive behavior (Albers & Grieve, 2007). The Bayley-III is only available in English. Accordingly, Williams, Sando, & Soles (2014) rated the Bayley-III as "inadequate" in its utility for children that do not speak English.

The Bayley-III was updated from the Bayley Scales of Infant Development, Second Edition (BSID-II), and was published in 2005. While the BSID-II used a combined mental developmental index (MDI), the Bayley-III has separated this into domains of cognition, language, and motor development (Yu et al., 2013). The cognitive portion consists of 91 items. The language portion is divided into receptive and expressive subparts, with each part containing 49 and 48 items, respectively. Studies have found that BSID-II and Bayley-III scores are highly correlated, and that the use of the language domain is effective in assessing early language development, with appropriate distinction of language from cognition (Yu et al., 2013; Albers & Grieve, 2007).

Administration and scoring. Examiners using Bayley-III are recommended to have a Master's degree or graduate level training, in addition to familiarity with administering child assessments. This is particularly important since the cognitive and language scales of the Bayley-III are entirely based on direct child assessments, which require the examiner to establish trusting relationships with toddlers. This also requires sufficient energy levels from children to participate in the Bayley-III (Albers & Grieve, 2007; Scattone, Raggio, & May, 2011).

Table 1. Detailed descriptions of the shortlisted instruments.

Name	Domains assessed & type of instrument	Age group	Reliability & Validity	Administration time & qualifications required	Instrument cost/ Conservative cost estimate for the BCHCP (components outlined in Appendix E)
Battelle Developmental Inventory, Second Edition (BDI-2)	Personal-social, adaptive, motor, communication, and cognitive ability Examiner-administered child assessment	Birth to 7 years, 11 months	Test-retest reliability >0.80 Inter-rater reliability >0.90 Cognitive and communication scores correlates with BSID-II's MDI (r = 0.61, 0.75, respectively) Communicative score correlates with PLS-4 total score (r = 0.72) Generally has moderate correlations with the other child cognitive and language instruments Lacking evidence on predictive validity	60-90 minutes for entire test; cognition & communication likely between 30 minutes to 1 hour (not including breaks for children between subscales); 15-20 minutes to score Some training is required for test administrators: general college-level training in measurement and statistical concepts	$1,282 USD for Complete Kit $79.20 USD for 15 additional record forms (individual pricing for other components on publisher website) Estimated total cost: $36,100
Bayley Scales of Infant and Toddler Development, Third Edition (Bayley-III)	Cognitive, motor, language, interaction, social-emotional, adaptive behaviour Examiner-administered child assessment for cognitive, motor and language domains Parent reports for interaction, social-emotional and adaptive behaviour domains	1 month to 3 years, 6 months	Test-retest reliability >0.80 Inter-rater reliability for cognition = 0.87–0.97; for language = 0.76–0.95 Cognitive and communication scores correlates with BSID-II's MDI (r = 0.62-0.78, 0.30–0.89, respectively) Correlations with other child cognitive and language instruments, including PLS-4, were all ≥0.51 Lacking evidence on predictive validity: BSID-II at 18 months development was predictive of development at 36 months (MDI, p<0.001)	Up to 90 minutes for entire test for children >13 months; cognitive and language likely between 30 minutes to 1 hour (not including breaks for children between subscales); 15-20 minutes to score Training is required for test administrators: Master's degree or graduate level training, as well as some familiarity with completing assessments and administering tests, or having a degree or license to practice in healthcare	$1,135 USD for Comprehensive Kit $122.25 USD for 25 additional cognition, language, & motor forms (individual pricing for other components on publisher website) Estimated total cost: $34,300

Language Development Survey (LDS)	Expressive vocabulary and word combinations, language delay screening Parent report	18 months to 2 years, 11 months	Test-retest reliability: 0.97 for vocabulary, 0.87 for mean phrase length Correlations with other instruments, including BSID-II's MDI, and PLS, range from 0.72-0.87; correlation with CDI-WS: 0.90 for vocabulary and mean length of phrases, 0.84-0.94 for lexical development In determining language delay, sensitivity is roughly 80%, specificity is roughly 85% Vocabulary scores higher in girls than boys; lower vocabulary scores in bilingual families Correlation between low SES and vocabulary scores = -0.14 Lacking evidence on predictive validity	10 minutes for parents to complete (can be mailed, and likely interviewed by professional); estimated to take 5 minutes to score Minimal training required for administrators: administrators typically score the completed forms and generate scores using the instrument materials	$55 USD for Preschool Manual plus Multicultural Supplement $30 for 50 LDS forms $30 for 50 LDS hand-scoring norm sheets Estimated total cost: $5,700
Language Use Inventory for Young Children (LUI)	Social pragmatic and expressive use of language Parent report	18 months to 3 years, 11 months	Test-retest reliability = 0.75-0.89 Internal validity of subscales range from 0.8-0.90 In determining language delay, sensitivity and specificity both = 95.9% Some evidence that SES and maternal education can bias results Limited evidence on predictive validity: low correlations were found between the LUI test at 24-29 months and the direct child language assessment at mean age of 5 years, 8 months (r = 0.29-0.33); sensitivity and specificity in predicting language delay were 0.81, and 0.93, respectively	20 minutes for parents to complete (can be accessed electronically, mailed, and likely interviewed by professional) Minimal training required for administrators: test is automatically scored by online software; manual scoring can be done within 5 minutes	Online: $3 CAD per assessment Hardcopy: $165 CAD for Starter Kit $150 CAD for manual Refill Pack (50 LUI & score sheets) Estimated total cost: $7,100

MacArthur-Bates Communicative Development Inventory (CDI)	Language and communication skills Parent report	8 to 3 years, 1 month CDI: Words and Sentences for use in children 16-30 months	Moderate to high correlations for vocabulary production at 2 years Test-retest reliability of vocabulary (Over 6 months) = 0.75-0.89; of grammatical complexity (over 6 months) = 0.59-0.61 Test-retest reliability (over 1 month) = 0.95 High correlations with other childhood language instruments In determining parent-reported language delay, sensitivity is 50-68%, specificity = 98% Scores higher in girls than boys Some evidence that SES and maternal education can bias results Limited evidence on predictive validity: mixed results of predictive validity from 2 years to school age, but generally favourable	20-40 minutes for parents to complete (can be interviewed by professional, and likely mailed) and 10-15 minutes to score Minimal training required for administrators: administrators typically score the completed forms and generate scores using the instrument materials	$59.95 USD for User's Guide and Technical Manual $25 USD for 25 Words and Sentences forms (individual pricing for other components on publisher website) Estimated total cost: $13,000
Preschool Language Scale, Fifth Edition (PLS-5)	Total language, auditory comprehension, expressive communication standard scores, growth scores, percentile ranks, language age equivalents Examiner-administered child assessment	Birth to 6 years, 11 months	Lacking peer-reviewed published results for PLS-5: non-peered review results: PLS-5 test-retest reliability = 0.86-0.95; inter-rater reliability was between 0.96-0.99 Peer reviewed: PLS-4 expressive communication inter-rater reliability = 0.99; content validity favourable for assessing and differentiating small differences in language development Some evidence that SES can bias results Lacking evidence on predictive validity	45-60 minutes for children >1 year (likely does not include breaks for children between subscales); 15-20 minutes to score Training is required for test administrators: Master's degree or graduate level training, as well as some familiarity with completing assessments and administering tests, or having a degree or license to practice in healthcare	$358.75 USD for Complete Kit with Manipulatives $162.50 USD for 50 additional PLS-5 record forms (individual pricing for other components on publisher website) Estimated total cost: $25,900

Administering the entirety of Bayley-III is estimated to take up to 90 minutes for children 13 months and older (Albers & Grieve, 2007). Including breaks for children between or during the subtests has yielded full administration times between 2.5 to 3 hours (Scattone, Raggio, & May, 2011). Although there is no peer-reviewed source on the time to complete select domains, it is estimated that the time required to complete the cognitive and language domains ranges between 30 minutes to 1 hour.

The scoring of the Bayley-III is based on a standardized US sample of 1700 children in 2000. Raw scores from the Bayley-III cognitive and language scales are converted to composite scores using this standardization sample. Within this sample, 10% were considered to have mental, physical, or behavioral issues in development, which aims to reflect the general population when generating Bayley-III test scores, and assists in identifying developmental delays. Normative data based on this sample are available in increments of 10 days to generate scores for children across the age range for the Bayley-III (Albers & Grieve, 2007).

Reliability and validity. The Bayley-III is recognized as a reliable instrument. Scale composite average reliability coefficients were 0.91 for the cognitive scale, and 0.93 for the language scale (Albers & Grieve, 2007). Test-retest reliability, with a mean retest time of 6 days, gave average coefficients of 0.80 or higher. The inter-rater reliability of the Bayley-III are regarded as excellent, with coefficients for the cognitive scale ranging from 0.87–0.97, and the language scale from 0.76–0.95 (Yu et al., 2013).

The Bayley-III has favourable concurrent validity with a number of previously validated child development instruments, including its previous version. The separated cognitive scores of the Bayley-III correlate well with the previous composite BSID-II MDI score (r = 0.62-0.78) (Albers & Grieve, 2007; Yu et al., 2013). Correlations for the language items ranged from 0.30–

0.89 (Yu et al., 2013). Correlations with other examiner-administered and parent-reported instruments, including the Wechsler Preschool and Primary Scale of Intelligence III, the Preschool Language Scale-4, and the Vineland Adaptive Behavior Scale (Appendix C) were all 0.51 or higher.

Evidence on Bayley-III's predictive validity was lacking for test administration at 24 months of age. However, since the Bayley-III is meant to be an improvement of the BSID-II, the predictive validity of BSID-II may be able to provide some indication of Bayley-III's predictive validity. Lung et al. (2009) reported on BSID-II's predictive validity for children from 6 to 36 months. Development at 18 months was found to be predictive of development at 36 months (MDI, $p<0.001$). However, this study did not explicitly assess whether administration at 24 months is predictive of development at 36 months or beyond. Given the limited evidence available, further research is warranted to draw conclusions on the predictive validity of Bayley-III (Johnson & Marlow, 2006).

Battelle Developmental Inventory, Second Edition (BDI-2)

Description. The BDI-2, published in 2005, assesses the domains of personal-social, adaptive, motor, communicative, and cognitive development in children from birth to 7 years, 11 months of age. The BDI-2 has been translated to Spanish, however, there is limited evidence on its accuracy and validity in assessing child development. In terms of using the BDI-2 for children that do not speak English, Williams, Sando, & Soles (2014) rated BDI-2 as "inadequate". Like the Bayley-III, BDI-2 aims to identify strengths and areas for growth in typically developing children and those affected by disabilities. The BDI-2 also assesses signs of developmental delays. The cognitive domain has a total of 105 items, while the language domain has 95 language items. Similar to the Bayley-III, the communication domain is divided into receptive

and expressive parts. The amount of items administered is dependent on the age of the child (Alfonso, Rentz, & Chung, 2010).

Administration and scoring. Compared to other comprehensive child development instruments, the BDI-2 requires lower examiner qualifications for test administration (Johnson & Marlow, 2006). College-level training, with formal education in general measurement and statistics are recommended to interpret BDI-2 results. Graduate-level research assistants have successfully administered BDI-2 for preschool children following a 1-day training session (Hallam et al., 2014).

Administration of the BDI-2 in its entirety requires 1 to 1.5 hours (Alfonso, Rentz, & Chung, 2010). Although research evidence is unavailable on administering only the cognitive and communication subscales, it is likely that these subscales would not take over an hour to complete. In contrast to the Bayley-III's reliance on direct examiner-child assessment, the BDI-2 presents 3 options of techniques that can be used during test administration: 1) direct assessment, 2) observation, and 3) interviewing the caregiver. These options are ranked to indicate the ideal method to assess each particular item in the BDI-2, while offering the remaining alternatives should the preferred method not be possible. Some judgment is required from the examiner to decide which methods are ideal in the test environment (Alfonso, Rentz, & Chung, 2010).

Scoring of the BDI-2 is based on a standardized sample of 2500 children that approximates the 2000 US census in age, sex, ethnicity, geography, and SES (based on mother's education level), and includes special group studies with children with disabilities (Johnson & Marlow, 2006; Macy, Bagnato, Macy, & Salaway, 2015). Normative data are available at increments of 3 months for children 24 months and older (Alfonso, Rentz, & Chung, 2010).

Reliability and validity. At 24 months of age, the BDI-2 is regarded as a reliable instrument (Alfonso, Rentz, & Chung, 2010). Test-retest reliability on children varying in sex, ethnicity, and SES, with a mean retest time as 8 days, gave good reliability coefficients for all domains ($r > 0.80$). Inter-rater reliability is also >0.90 (Macy et al., 2015).

The BDI-2 has been validated against a wide range of examiner-administered and parent-reported instruments, including the BSID-II, the Preschool Language Scale (PLS), and the Vineland Social Emotional Early Childhood Scales (Appendix C) (Bliss, 2007). In general, the BDI-2 has acceptable correlations with these instruments ($r > 0.70$) (Bliss, 2007; Alfonso, Rentz, & Chung, 2010; Macy et al., 2015). These comparisons have provided further support of using the BDI-2 to assess child development, with the BDI-2 cognitive and communication scores correlating well with BSID-II's MDI ($r = 0.61, 0.75$, respectively). The correlation between the BDI-2 communicative score and the PLS-4 total score was 0.72 (Alfonso, Rentz, & Chung, 2010). The current review did not find results on BDI-2's predictive validity.

Preschool Language Scale, Fifth Edition (PLS-5)

Description. The PLS-5, published in 2011, is used to assess language development in children, providing insight on strengths and weaknesses in receptive and expressive language use (Leaders, 2013). Like the BDI-2, PLS-5 is designed for children from birth to 7 years, 11 months, and is also available in Spanish. Unlike the Bayley-III and BDI-2, there are only two scales within the PLS-5: auditory comprehension and expressive communication. Given that the PLS-5 is designed to assess language specifically, the test items within these scales provide an in-depth assessment of language development in syntax, expression, grammar, and sentence use (Leaders Project, 2013). In addition to this, there are also 3 optional subscales: the Language Sample Checklist, the Articulation Screener, and the Home Communication Questionnaire.

These 3 supplemental materials are designed to provide additional support to the PLS-5 items used to interpret the total language score. However, there is limited scientific evidence on the use of these supplemental materials, and one reviewer criticized the validity and value of these materials (Cleave, 2006).

Administration and scoring. Like Bayley-III, PLS-5 examiners are recommended to have Master's degrees or graduate level training, as well as familiarity with completing diagnostic assessments and administering tests. Similar to the Bayley-III and BDI-2, it is estimated that the PLS-5 can be administered within 45-60 minutes for children over 1 year old. However, it is unclear whether this timeframe includes times for breaks (Leaders, 2013).

The PLS-5 has a standardization sample of 1400 children representative of the 2008 US census, in terms of age, sex, ethnicity, geography, and SES (based on mother's education level) (Leaders Project, 2013). Within this sample, 3% spoke languages other than English at home. Normative data are available at 6 month intervals for children over 1 year old.

Reliability and validity. The systematic search yielded no peer-reviewed results on the validity and reliability of PLS-5. One non-peer-reviewed report on the PLS-5 found the test-retest reliability to be 0.86-0.95 with a retest interval of 3-28 days. The inter-rater reliability was between 0.96-0.99. However, this same test review criticized the samples used to generate these coefficients, as they may be too small to establish sufficient statistical power (195 and 54 children, respectively).

Without peer-reviewed results on the PLS-5, evidence on validity can be loosely drawn from studies on the PLS-4. The PLS-4 is a widely used instrument that has been standardized and shown to be valid and reliable for use in typically developing and clinical preschool children (Limbos & Joyce, 2011; Zimmerman & Castilleja, 2005). The PLS-5 was created as an update to

the PLS-4, particularly to reflect new research on language development. It is presumed that the item changes found in the PLS-5 will help the instrument better assess language development according to current research and practices. The PLS-4 has an excellent inter-rater reliability of 0.99 for the expressive communication portion of the instrument (Zimmerman & Castilleja, 2005). PLS-4 was validated against the Denver-II (Appendix C), and the instrument also has favourable validity in assessing and differentiating small differences in language development (r >0.70). The current review did not find results on PLS-5's predictive validity.

Some evidence also exists to suggest that the PLS-5 scores can be biased by SES. This has been attributed to children's vocabulary exposure. It has been shown that families with higher SES use more extensive vocabulary than families with lower SES. A child living in a lower SES environment may then be less able to answer PLS-5 questions that are reliant to vocabulary exposure. Hence, there runs a risk that these children may be identified as at-risk of, or having a language disorder, when in reality there is no issue in the development process, but rather a deficit in experience (Leaders, 2013).

Parent-reported instruments for language

MacArthur-Bates Child Communicative Inventories (CDI)

Description. The CDI relies on parent-reported information on their children's vocabulary and language development (Law & Roy, 2008). Due to its popularity, many researchers and clinicians have adapted the three scales of the CDI, and hence many variations exist, including shorter forms of the CDI. The short forms, however, have been less studied (Deniz Can et al., 2012). The CDI has been translated for use in many different languages – 42 versions were available as of 2008, including a sign language version (Law & Roy, 2008).

It consists of 3 scales that can be used independently for different age ranges, with one of them being used for children between 16-30 months of age: the CDI Words and Sentences (CDI-WS) (Feldman et al., 2005). The other 2 scales are designed for use in children 8-18 months, and children 30-37 months. CDI-WS is a checklist of 680 words in 22 categories used to assess language production, including grammar and syntactic development (Rescorla et al., 2005). CDI-WS also requests parents to report on the best three sentences they have heard from their children. It should be noted that the CDI-WS, like the PLS-5, does not assess language comprehension, which is attributed to some evidence of limited validity of parental reports in this area for 24 month old children (Law & Roy, 2008).

Administration and scoring. Without direct child assessment, limited training is required to administer the CDI. Each scale of the CDI is estimated to take 20-40 minutes for parents to complete. Scoring typically takes 10-15 minutes for each scale to generate a language score for the child. All CDI scales can be mailed, and if necessary, can be read to parents to aid completion (Feldman et al., 2005).

The standardization samples for the CDI (n=1803), and CDI-WS (n=1130) have been criticized for not including sufficient amounts of minority groups to be representative of the US population. This poses as an additional problem to relying on parental reports when administering the CDI to minority populations (Law & Roy, 2008; Rescorla & Alley, 2001). The current review did not find information on normative data increments that are available with scoring the CDI.

Reliability and validity. The CDI reliability coefficients are slightly higher in older children (i.e. 25 months, as compared to 19 months), likely due to further grammatical development (Law & Roy, 2008). Reliability of the CDI's assessment of vocabulary is

considered acceptable to good over an approximate retest interval of 6 months (r = 0.75-0.81), whereas assessment of grammatical complexity was slightly lower (r = 0.59-0.61). Test-retest reliability over 1 month was found to be excellent (r = 0.95) (Rescorla et al., 2005).

A review of the CDI's concurrent validity found moderate to high correlations for vocabulary production in children at 24 months of age (Law & Roy, 2008; Deniz Can et al., 2012). High correlations have been found between CDI-WS and other childhood language instruments (Rescorla et al., 2005). Because the CDI has 3 scales for subgroups within the overall age range, it is also of interest to understand the correlations between these scales. Unacceptably low correlations have been found between the score of the CDI scale used at 12 months of age and the CDI-WS administered at 24 months (r = 0.18-0.39). This has been attributed to variability in the judgment of language for children at these developmental milestones. Mothers may overestimate language abilities at 12 months due to lower expectations at this point in development. It has been suggested that mothers are more able to accurately judge their children's language at 24 months. Hence, caution should be taken when using CDI scores to assess child language development over time, particularly if using the instrument to assess treatment effectiveness (Feldman et al., 2000).

There were few studies exploring the predictive validity of the CDI. In looking at outcomes of children in the subsequent 8-21 months after CDI administration, correlations were found to be roughly 0.50 (Law & Roy, 2008). Correlations were higher for vocabulary scores rather than grammar scores. One study found that CDI scores at 6 months were predictive of scores at 24 months. Another study found total vocabulary size at 24 months was predictive of language skill in grade 5 (Deniz Can et al., 2012). The CDI-WS administered at 24 months was concluded by Feldman et al. (2005) to be fair to good in predicting language skills at 36 months.

To screen for language delay, the CDI-WS's sensitivity was found to range from 50-68%, and specificity was 98% (Law & Roy, 2008). This can lead to false positives in the prediction and identification of language delay (Pesco & O'Neill, 2011). At 24 months of age, girls were found to score higher than boys on the CDI (Feldman et al., 2005). It has been suggested that this is reflective of the earlier language development of girls (Lung et al., 2009).

Some evidence also suggests that parent-reporting can be affected by sampling characteristics, such as SES (Rescorla et al. 2005; Law & Roy, 2008). For example, the use of the CDI in New Zealand and the US found that parents with lower education over-reported their children's vocabulary development. However, in a sample of African-American children of 30 months, the opposite effect was observed. A possible strategy to decrease bias is to have multiple reporters on the same child – however, evidence for this is limited. Hence, reviewers have recommended taking caution in using the CDI to assess language deficits and comparing children coming from differing SES backgrounds (Law & Roy, 2008).

Language Development Survey (LDS)

Description. The LDS is a vocabulary checklist used to assess language development in children, and is most widely used when children are 24 months old, although the instrument is designed for use in children from 18-35 months (Rescorla, Ratner, Jusczyk, & Jusczyk, 2005). The LDS checklist has been translated to several other languages, and consists of 310 words organized into 14 categories (Rescorla & Achenbach, 2002). The LDS is often used for screening language delays. Like the CDI, the LDS also asks parents to report on three of their children's best sentences (Rescorla et al., 2005).

Administration and scoring. Without the need of professional examiners, the LDS has been used in many different contexts, including clinics, daycares, home interviews, mail, and

inner-city environments (Rescorla et al., 2005). Parents can complete the LDS within 10 minutes, and requires parents to have fifth-grade reading skills. The LDS provides reference norms from 18 to 35 months (Rescorla & Achenbach, 2002). At 24 months, children with results at 10% of the referenced population, or having fewer than 50 words, are identified as at-risk of, or having delays (Rescorla et al., 2005). It shares its standardization sample with a child behaviour assessment tool, the Child Behavior Checklist for Ages 1½-5, which included 278 children with SES, ethnic, and language diversities (Rescorla & Achenbach, 2002).

Reliability and validity. With a mean retest interval of 23 days in 66 children, the LDS was found to have good to excellent test-retest reliability, with coefficients of 0.97 for vocabulary, and 0.87 for mean phrase length (Rescorla et al., 2005). The LDS has been validated against numerous tests, including the BSID-II MDI and the expressive vocabulary portion of the PLS ($r = 0.72$-0.87). The LDS has also been validated against the CDI-WS, with correlations above 0.90 for the scores of vocabulary and mean length of phrases (Rescorla et al., 2005). The LDS was found to be inferior to the CDI-WS in assessing children above the 90th percentile in vocabulary development. This validation study, however, was based largely on White, middle-upper class families, and may be subjected to selection bias. No evidence was found within this review on LDS's predictive validity. In determining language delay, sensitivity of the LDS is roughly 80%, while specificity is roughly 85%. Similar to the CDI, LDS scores have been found to be higher in girls than boys at 24 months (Rescorla et al., 2005; Rescorla & Alley, 2001).

When used in bilingual families, the LDS yielded lower vocabulary scores than monolingual families, but found no significant differences in the mean length of phrases used by children. It has been suggested that this is reflective of bilingual children using equivalent words in their other language(s), resulting in the use of bilingual phrases and parental underreporting of

the full set of vocabulary they know. LDS scores have also been found to vary significantly in different ethnic groups, and this may be attributed to differences in parent-reporting, which may also be affected by SES. The relationship between lower SES and LDS vocabulary scores has been correlated at -0.14 (Rescorla et al., 2005).

Language Use Inventory for Young Children (LUI)

Description. Compared to the other shortlisted instruments, the LUI is a relatively new instrument published in 2002 (O'Neill, 2007). The LUI is designed to assess language development in children aged 18-47 months. In particular, the LUI aims to assess children's pragmatic and expressive language competence, which contrasts to the vocabulary and grammatical focus of other instruments, including PLS-5, CDI, and LDS. It is noted, however, that even with the focus on pragmatics, semantics and syntax are still assessed in parts of the 14 LUI subscales, since these domains are correlated with one another (Pesco & O'Neill, 2011). To assess children's social use of language, three main domains are assessed: communication with gestures, communication with words, and the use of longer sentences, with a total of 180 items (O'Neill, 2007).

Administration and scoring. Like other parent-reported instruments, the LUI does not require highly trained examiners, and takes parents 20-30 minutes to complete. Two of the subscales are open-ended questions. The norming sample is 3563 Canadian children, with diversities in income, parent education, family structure, and cultural and ethnic backgrounds. Norms are available at one month intervals, and are also available by gender, which can help alleviate the issue of gender differences present in CDI and LDS scores. The LUI is available in both hardcopy and electronic versions. The electronic version can be disseminated using an online platform, as well as on a mobile device brought to parents. Scoring is automated with the

software. Scoring of the hardcopy LUI can be manually completed by the administrator within 5 minutes (Pesco & O'Neill, 2011).

Reliability and validity. The LUI appears to be a valid instrument that is sensitive to the changes in children's language development from 18-47 months. Test-retest reliability over a 4-week interval was found to range from 0.75-0.89. The validity of the LUI subscales is good, with the majority of values ranging from 0.80-0.90 (O'Neill, 2007). Because the LUI is relatively new, there were few studies examining its predictive validity. After administering the LUI at 24-29 months of age, low correlations were found between LUI scores and examiner-administered language assessment scores found at follow-up ($r = 0.29$-0.33). The mean follow-up age was of 5 years, 8 months. However, the sensitivity and specificity in predicting language delay over this time period were 0.81, and 0.93, respectively (Pesco & O'Neill, 2011). Another study found the sensitivity and specificity of the LUI in determining language delay to both be 95.9% (O'Neill, 2007). Similar to other parent-reported instruments, some evidence suggests that SES and maternal education can bias the results of the LUI.

Discussion

The BCHCP is a rigourous research project that aims to collect extensive data to understand the potential effects of NFP on maternal and child outcomes in BC. In doing so, many measures and instruments are used throughout the trial, with the majority of the child measures being administered at the endpoint of NFP when the children are 24 months of age, including child cognitive and language development. While this provides comprehensive data, important considerations need to be taken to ensure the feasibility of the instruments (i.e., in budgeting, training, and administration), while maintaining data quality to meet the objectives of the

BCHCP. The chosen instrument(s) must be reliable, valid, and appropriate for young, first-time mothers and their children that may be facing social and economic disadvantage.

There are several strategies to consider to decrease the resource burden for the BCHCP in the assessment of children at 24 months of age. If administering a comprehensive cognition and language instrument (i.e., Bayley-III or BDI-2) is not possible for the entire BCHCP sample, an option worth exploring would be using a cognitive and language instrument (i.e., Bayley-III, BDI-2) for a smaller subsample of the BCHCP, while administering a parent-reported language instrument (i.e., CDI, LDS, or LUI) for a larger subsample, if not the entire sample. If sufficient power can be maintained, the BCHCP could also only administer the Bayley-III or BDI-2, without using a parent-reported language instrument. Given some evidence that cognition is a precursor to language development, in addition to the concurrent validity of the parent-reported language instruments shortlisted in this review, assessing language development for a larger sample in addition to the Bayley-III or BDI-2 subsample can provide supplemental information that is indicative of cognitive development, even though it would not be the most ideal and direct indicator. There are limitations in generalizing these results to draw definitive conclusions on cognitive development, since the overlap between cognition and language at 24 months of age has not been quantified (Siegel, 1981). Nonetheless, having some data on language development for these participants holds more benefit than having no data on these participants. A larger sample size would be beneficial in identifying NFP effects on child development. Depending on the BCHCP's resource considerations, the instruments and the extent that they are used can be justified in a multitude of ways, but it is not recommended to forgo a cognitive measure (i.e., Bayley-III, BDI-2) entirely if the BCHCP aims to draw conclusions on cognitive and language development.

An additional consideration is the importance of ensuring that at the very least, a sufficient portion of the BCHCP sample can communicate in English. This is necessary to maintain statistical power to conclude any significant differences due to the effects of NFP, since the Bayley-III is only available in English, and the BDI-2 has only been translated to Spanish. This would also apply in the scenario where parent-reported instruments are administered, since instruments may falsely identify language errors due to ESL as language delays or impairments, even if the children do not experience any issues in their other language(s). Although the BCHCP's eligibility criteria requires the mothers' abilities to converse in English, it will be important to confirm that the children at 24 months of age have been communicating in English in the home before administering the shortlisted instruments recommended in this current review.

The Bayley-III, BDI-2, and PLS-5 are expensive and time-consuming instruments. Although they are play-based to induce children's participation, the validity of the results is compromised if children refuse to interact with the examiner, particularly with the Bayley-III. For children at 24 months of age, it is likely that examiners will need to devote time for rest breaks, or be prepared to reschedule the assessment entirely if children are not cooperative. Further complications may arise if these tests are completed over a span of visits, as children's performance can be influenced by a multitude of factors, including variability in the testing environment and the developmental state of children between visits. It is hence recommended to complete the assessments in their entirety within one session.

Overall, if administered properly, the BDI-2 and Bayley-III are comparable in reliability and validity, and can be advantageous for several reasons. Both instruments were rated "good" for use in children that speak English, and are highly regarded due to their recent standardization samples (Williams, Sando, & Soles, 2014).

The advantages of the Bayley-III lie in its popularity, its extensive body of evidence, and its excellence in providing objective assessments on child development (Limbos & Joyce, 2011). However, the trade-off to this is that the Bayley-III is a rather rigid test, with high standards in training and administration. Qualifications to conduct the Bayley-III are high, and more extensive training is required before one can successfully administer the instrument. Further, because the Bayley-III is an instrument that directly assesses children, the tasks may be overwhelming for young children (Scattone, Raggio, & May, 2011). Experienced Bayley-III examiners will ease administration, particularly in developing rapport with children, since the instrument will not yield valid results without children's full participation.

The BDI-2 is regarded as a comprehensive and user-friendly instrument. Compared to the Bayley-III, less examiner training is required. Test administration is also more flexible, allowing the use of direct assessment techniques, observations, and parent interviews. Adequate examiner training and practice to understand which technique is ideal under different situations can yield valid and reliable results (Alfonso, Rentz, & Chung, 2010). While the BDI-2 may be beneficial in environments with difficulty in direct assessment, the use of examiner observations and parental interviews opens up the opportunity for examiner bias (Hallam et al., 2014). However, this bias can be mediated with sufficient training, leading BDI-2 to be comparable to Bayley-III as highly valid and reliable instruments (Williams, Sando, & Soles, 2014).

For the BCHCP, the Bayley-III would undoubtedly be the most comparable instrument to previous NFP research (Appendix C), in addition to its reliability and validity to assess child development in the BCHCP. However, planning and budgeting is critical to ensure the use of trained, skilled scientific field interviewers to establish rapport with young children and deliver the specific child assessment techniques. Should this not be possible, the BDI-2 presents as a

worthy substitute to the Bayley-III. The BDI-2 also has the advantage of various administration techniques, as well as the opportunity for comparisons in longitudinal follow-up. Compared to the upper age range of the Bayley-III at 3 years, 6 months, the BDI-2 can be used up until 7 years, 11 months. This can be particularly useful if the BCHCP aims to assess child development for the BCHCP sample when the children reach school-age, especially given the poor predictive validity of child development instruments. To draw powerful conclusions on longitudinal outcomes, further assessments must be made, and the BDI-2 provides the option for being repeated for direct comparisons with previous data.

The distinct language and communication domains of the Bayley-III and BDI-2 also eliminate the need for an additional test for language development, particularly due to their favourable correlations with the PLS-4. If one of the Bayley-III or BDI-2 were to be used at 24 months in the BCHCP, administering the PLS-5, or any of the parent-reported language instruments to the same participants would add limited value.

There are concerns surrounding the performance of the PLS-5. Without sufficient published articles on the PLS-5, it is difficult to ascertain its validity and reliability. Although it is likely to be an improvement of the PLS-4, further research would be ideal to support PLS-5's use in assessing intervention effectiveness on child language development. It is worth mentioning that the PLS-4 is a valid and reliable instrument, and is one of the few language instruments not reliant on parental reports. Although the norming sample may be older, the PLS-4 may still be of value to the BCHCP. However, it would be more economical to use one of Bayley-III or BDI-2 in place of the PLS-5, as the former 2 tests can also directly assess cognition in addition to language within a similar timeframe.

In contrast to the resource-intensive examiner-administered instruments, parent-reported instruments are quick and can yield information that may not be present in testing environments (Rescorla & Alley, 2001). It is important to note, however, that the shortlisted parent-reported instruments in this review are intended to assess language development. Using these parent-reported instruments alone would not provide sufficient information to draw meaningful conclusions on cognitive development. These instruments can be best used to supplement the findings of the Bayley-III or BDI-2 in the BCHCP.

The CDI, LDS, and LUI can be completed without an in-person visit. The CDI and LDS have been completed via mail in previous research studies (Rescorla et al., 2005; Feldman et al., 2005), and can likely be completed via telephone, as a method to increase response rates. In addition to hardcopies, the LUI has also been designed for online completion, which can ease efficiency in data collection. Further, the LDS is designed with instructions at a fifth-grade reading level, which works to the benefit of the BCHCP sample with a proportion of participants being eligible due to lower levels of education. However, with the use of scientific field interviewers in the BCHCP, in addition to the checklist portions within these parent-reported instruments, it is unlikely that the CDI and LUI instructions will pose as difficulties for the BCHCP participants.

Of the 3 shortlisted parent-reported language instruments, the CDI is the most in-depth, and can provide more detailed measures than the LDS and LUI. However, despite its popular use, the sensitivity of CDI was found to be substantially lower than the LDS and LUI in determining language delay. This finding may be partially due to the larger body of evidence behind CDI, whereas the majority of studies found for the LDS and LUI were from the publishers of the instruments. The CDI and the LDS remain as popular instruments to assess

language development in infants, due to its easy administration and cost-effectiveness (Rescorla & Achenbach, 2002; Law & Roy 2008; Skarakis-Doyle, Campbell, & Dempsey, 2009). Because of its recent publication, there are fewer published articles on the LUI, but it appears to be a promising instrument. A defining aspect of the LUI is that it is a Canadian instrument, using a standardized Canadian sample for referencing scores. The LUI also provides norms by gender, which attempts to resolve the gender discrepancy present in both CDI and LDS.

The most fitting parent-reported language instrument for the BCHCP would depend on what the trial seeks to evaluate within language development. Should more detailed and technical information (i.e. vocabulary, syntax, grammar) on language development be required, the CDI would be the ideal instrument. Compared to the LDS, the CDI also has more favourable evidence on its predictive validity. Should the aim be screening for language delay, both the LDS and LUI are valid and reliable, and have been shown to perform better than the CDI in identifying delays, in addition to their shorter administration times. The LDS is favourable for its convenience and correlations with PLS and BSID, while the LUI is favourable for its Canadian norms and some evidence for predictive validity. It is important to note that the LUI is focused on pragmatic language, and hence if the BCHCP is seeking detailed information on expressive and receptive language rather than the social use of language, the LUI would not be ideal.

Although the CDI has been used in evaluations of home visitation programs documented by HomVEE, it has not generated conclusions of significant program effects (Appendix B). Although reliability and validity are favourable, without evidence of their use in other home visitation programs within this review, it is also unclear whether the PLS-5, LDS, and LUI would perform well for the BCHCP in terms of being sensitive to small changes due to NFP in a community sample. Taking into consideration the body of scientific evidence, the Bayley-III has

the highest potential of producing reliable and valid results for the BCHCP. If there are training and administration concerns, the BDI-2 would serve as an appropriate substitute.

Limitations and future direction

One of the objectives of this review was to explore the predictive validity of the shortlisted measures. Unfortunately, the search yielded very limited articles in this area. Two main reasons are proposed for this, with the first being that some of the shortlisted instruments are relatively new – there simply has not been enough time to study these instruments' predictive validities. The second reason is that developmental tests for young children have historically been found to perform poorly in predicting later outcomes, due to the high probability of measurement error in working with infants and toddlers (Johnson & Marlow, 2006). Predictive validity is also especially poor for typically developing children, since there is a wide variation of development patterns in early childhood, in conjunction with the complex influences of environmental, social, and medical factors (Pesco & O'Neill, 2011). It is likely that developmental testing for clinical populations, and the general population of children over 24 months old have a higher degree of predictive validity – however, this was not the topic of this current review. Without strong predictive validity in infant testing, longitudinal studies are ideal in improving our understanding of early childhood factors on cognition and language development. Despite these limitations, early childhood assessment remain an important practice to indicate potential needs for early intervention for developmental delays or disorders, as well as to assess the multitude of variables that may affect child development.

There also appears to be limited evidence and research completed for child cognition and language instruments for use in ESL populations. Although the majority of the shortlisted instruments are available in other languages (BDI-2, PLS-5, LDS, and CDI; LUI - ongoing

research for validation in other languages), it is recommended that test interpreters be aware that the norm-referenced scoring for these instruments are largely based on English-speaking populations, and hence scoring may not be as accurate for populations that are not fluent in English. For children not speaking English, cognitive testing options include using nonverbal instruments, or using one as an additional test to the BDI-2 or Bayley-III (Williams, Sando, & Soles, 2014). In parent-reported instruments, urging parents to also report on language-equivalent vocabulary, sentence and general communication in all of their children's languages may be beneficial in addressing the shortcomings of these instruments in ESL contexts.

Furthermore, low SES has generally been correlated with lower test scores. This was expected, since lower income and parental education have been found to negatively impact child cognitive and language development (Johnson & Marlow, 2006). In parent-reported language measures, extra caution should be taken, as low SES has been correlated with parental-over- and under-estimation of child language development (Rescorla et al. 2005; Law & Roy, 2008). Some research has recommended that standardization samples be created for various SES and ethnic variations (Law & Roy, 2008). However, the benefits of this would depend on the purpose of the developmental tests. If the aim is to assess child development under a certain environment as compared to the general population, then the current standardization samples of the shortlisted instruments should suffice, as they are generally representative of the national population (Rescorla & Achenbach, 2002).

Last, this review utilized four search engines, which leaves the possibility that some relevant articles may have been missed. Although sources were peer-reviewed, a considerable amount of articles found also authored by the creators of the instruments, which opens up the possibility of bias and vested interests in the outcomes reported were. Future research by

different authors, as well as longitudinal studies would be beneficial in more objective assessments of the validity and reliability of the instruments reported in this review.

Conclusion

It is widely accepted that child cognitive and language development indicators are effective in assessing the growth and developmental trajectories of children. However, the means of collecting this information can present as a challenge. No single child measurement instrument can be ideal under all contexts, particularly for the complex domains of children's cognition and language. Going forward, researchers will likely encounter deliberations of the trade-offs in using examiner-administered or parent-reported instruments, as well as test lengths, examiner training, testing environments, and demographic variables. However, with a thorough understanding of how these factors can be addressed in the interpretation of results, cognitive and language instruments can provide valid, reliable, and useful information to understand and assess child development.

Critical reflection

My capstone experience has deepened my understanding of the rigour behind public health research. When I enrolled at Simon Fraser University, I aimed to pursue my interests in child health while developing my skills in the principles and competencies of population health. Through the wide-ranging coursework in the Master of Public Health program, I was able to locate my passion to conduct research that would improve the lives of children and their families. My coursework in biostatistics, epidemiology, and child health policy were integral to the completion of this capstone project. The support that I received from the BC Healthy Connections Project (BCHCP) study team at the Children's Health Policy Centre at SFU has also provided me with the opportunity to apply the knowledge from the program, and gain valuable

experience in public health practice. I was able to critically analyze the published articles found in my literature search, and understand the implications of reliability and validity of instruments used in a large randomized controlled trial. I was also able to develop an appreciation for the considerations necessary to make sound methodology decisions in research.

In this systematic review I explored childhood cognition and language measurement instruments. Although I have worked with young children in the past, I have not had extensive formal training in child development. Completing this systematic review has not only taught me the main concepts of child cognitive and language development, but also the importance of designing a detailed plan before beginning a systematic review. This project served as a stark reminder that organization and clear objectives are critical in the creation of a successful product. In the future, I will take steps to ensure that I have a specific search strategy and eligibility criteria, in addition to a refined search topic before beginning the literature search. Aside from the technical aspects of conducting a systematic review, I also realized the positive impact that personal interest makes in overcoming the difficult challenges of a project. Knowing that my work is going to inform the BCHCP study team and their research decisions was the largest motivation for me to complete the project. Going forward, I believe it would be worthwhile to ensure that I maintain a clear perspective of the purpose of my work, and why I choose to do it.

I am delighted to have the opportunity to continue working with the BCHCP following the completion of this project and my MPH degree. I believe it will be an incredibly rewarding experience to work with a supportive team that is highly motivated in creating change. Working on this project has reinforced my passion of using research to inform health policy, and I hope to continue my learning in public health to work towards my goal of conducting research and translating research knowledge to help better the lives of children and families.

References

Achenbach System of Empirically Based Assessment. The Language Development Survey (LDS). Retrieved from http://www.aseba.org/research/language.html

Albers, C.A., Grieve, A.J. (2007). Test reviews – Bayley Scales of Infant and Toddler Development – Third Edition. *Journal of Psychoeducational Assessment, 25*(2), 180-198.

Alfonso, V.C., Rentz, E.A., & Chung, S. (2010). Review of the Battelle Developmental Inventory, Second Edition. *Journal of Early Childhood and Infant Psychology, 6*, 21-40.

Bliss, S.L. (2007). Test Reviews – Battelle Developmental Inventory − Second Edition. *Journal of Psychoeducational Assessment, 25*(4), 409-415.

Bradley-Johnson, S. (2001). Cognitive assessment for the youngest children: A critical review of tests. *Journal of Psychoeducational Assessment, 19*, 19-44.

Brookes Publishing. ASQ-3. Retrieved from http://www.brookespublishing.com/resource-center/screening-and-assessment/asq/asq-3/

Brookes Publishing. CDI. Retrieved from http://www.brookespublishing.com/resource-center/screening-and-assessment/cdi/

Caldera, D., Burrell, L., Rodriguez, K., Crowne, S. S., Rohde, C., & Duggan, A. (2007). Impact of a statewide home visiting program on parenting and on child health and development. *Child Abuse & Neglect, 31*(8), 829–852.

Curriculum Associates. BRIGANCE Early Childhood Screens III. Retrieved from http://www.curriculumassociates.com/products/detail.aspx?title=BrigEC-Screens3

Deniz Can, D., Ginsburg-Block, M., Golinkoff, R.M., & Hirsh-Pasek, K. (2013). A long-term predictive validity study: can the CDI Short Form be used to predict language and early literacy skills four years later? *Journal of Child Language, 40*(4), 821-835.

Denver Developmental Materials Inc. Denver II. Retrieved from http:// http://denverii.com/denverii/

Drotar, D., Robinson, J., Jeavons, L., & Lester Kirchner, H. (2009). A randomized, controlled evaluation of early intervention: The Born to Learn curriculum. *Child: Care, Health & Development, 35*(5), 643–649.

Feldman, H.M., Dale, P.S., Campbell, T.F., Colborn, D.K., Kurs-Lasky, M., Paradise, J.L., & Rockette, H.E. (2005). Concurrent and predictive validity of parent reports of child language at ages 2 and 3 years. *Child Development, 76*(4), 856-868.

Feldman, H.M., Dollaghan, C.A., Campbell, T.F., Kurs-Lasky, M., Janosky, J.E., & Paradise, J.L. (2000). Measurement properties of the MacArthur Communicative Development Inventories at ages one and two years. *Child Development, 71*(2), 310-322.

Fergusson, D. M., Grant, H., Horwood, L. J., & Ridder, E. M. (2005). Randomized trial of the Early Start program of home visitation. *Pediatrics, 116*(6), e803-e809.

Frances Page Glascoe. PEDS. Retrieved from http://www.pedstest.com/default.aspx

GL Assessment. Schedule of Growing Skills. Retrieved from http://www.gl-assessment.co.uk/products/schedule-growing-skills

Hallam, R.A., Lyons, A.N., Pretti-Frontczak, K., & Grisham-Brown, J. (2014). Comparing apples and oranges: The mismeasurement of young children through the mismatch of assessment purpose and the interpretation of results. *Topics in Early Childhood Special Education, 34*(2), 106-115.

Hamilton, S. (2006). Screening for developmental delay: reliable, easy-to-use tools. *Journal of Family Practice, 55*(5), 415-22.

Hogrefe. Griffiths Mental Development Scales - Revised: Birth to 2 years (GMDS 0-2). Retrieved from http://www.hogrefe.co.uk/gmds-0-2.html

Johnson, S., & Marlow, N. (2006). Developmental screen or developmental testing? *Early Human Development, 82*, 173-183.

Johnston, B. D., Huebner, C. E., Anderson, M. L., Tyll, L. T., & Thompson, R. S. (2006). Healthy Steps in an integrated delivery system: Child and parent outcomes at 30 months. *Archives of Pediatrics & Adolescent Medicine, 160*(8), 793-800.

King, T., Rosenberg, L., Fuddy, L., McFarlane, E., Sia, C., & Duggan, A. (2005). Prevalence and early identification of language delays among at-risk three year olds. *Journal of Developmental & Behavioral Pediatrics, 26*(4), 293-303.

Kitzman, H., Olds, D.L., Henderson Jr., C.R., Hanks, C., Cole, R., Tatelbaum, R., … Barnard, K. (1997). Effect of prenatal and infancy home visitation by nurses on pregnancy outcomes, childhood injuries, and repeated childbearing: A randomized controlled trial. *JAMA, 278*(8), 644-652.

Knowledge in Development. Language Use Inventory. Retrieved from https://languageuseinventory.com/

Landry, S. H., Smith, K. E., Swank, P. R., & Guttentag, C. (2008). A responsive parenting intervention: The optimal timing across early childhood for impacting maternal behaviors and child outcomes. *Developmental Psychology, 44*(5), 1335-1353.

Law, J., & Roy, P. (2008). Parental report of infant language skills: A review of the development and application of the communicative development inventories. *Child and Adolescent Mental Health, 13*(4), 198-206.

Leaders Project. (2013). *Test Review: Preschool Language Scales- Fifth Edition (PLS-5).* Retrieved from http://leadersproject.org/sites/default/files/PLS5-English-finaldraft.pdf

Limbos, M.M, & Joyce, D.P. (2011). Comparison of the ASQ and PEDS in screening for developmental delay in children presenting for primary care. *J Dev Behav Pediatr 32*, 499-511.

Love, J. M., Kisker, E. E., Ross, C., Raikes, H., Constantine, J., Boller, K., et al. (2005). The effectiveness of Early Head Start for 3-year-old children and their parents: Lessons for policy and programs. *Developmental Psychology, 41*(6), 885-901.

Love, J., Kisker, E., Ross, C., Schochet, P., Brooks-Gunn, J., Boller, K., et al. (2001). *Building their futures: How Early Head Start programs are enhancing the lives of infants and toddlers in low-income families. Summary report.* Report to Commissioner's Office of Research and Evaluation, Head Start Bureau, Administration on Children, Youth and Families, and Department of Health and Human Services. Princeton, NJ: Mathematica Policy Research.

Lowell, D. I., Carter, A. S., Godoy, L., Paulicin, B., & Briggs-Gowan, M. J. (2011). A randomized controlled trial of Child FIRST: A comprehensive home-based intervention translating research into early childhood practice. *Child Development, 82*(1), 193-208

Lung, F.W., Shu, B.C., Chiang, T.L., Chen, P.F., & Lin, L.L. (2009). Predictive validity of Bayley scale in language development of children at 6–36 months. *Pediatrics International, 51*, 666-669.

MacNamara, J. (1972). Cognitive basis of language learning in infants. *Psychological Review, 79*(1), 1-13.

Macy, M., Bagnato, S.J., Macy, R.S., Salaway, J. (2015). Conventional tests and testing for early intervention eligibility: Is there an evidence base? *Infants and Young Children, 28*(2), 182-204.

Madden, J., O'Hara, J., & Levenstein, P. (1984). Home again: Effects of the Mother-Child Home Program on mother and child. *Child Development, 55*(2), 636–647.

Mejdoubi, J., van den Heijkant, S., Struijf, E., van Leerdam, F., HiraSing, R., & Crijnen, A. (2011). Addressing risk factors for child abuse among high risk pregnant women: design of a randomised controlled trial of the nurse family partnership in Dutch preventive health care. *BMC Public Health, 11*, 823-831.

National Institute of Child Health and Human Development Early Child Care Research Network. (2000). The relation of child care to cognitive and language development. *Child Development, 71*(4), 960-980.

Olds, D.L., Henderson Jr, C.R., Tatelbaum, R., & Chamberlin, R. (1986). Improving the delivery of prenatal care and outcomes of pregnancy: A randomized trial of nurse home visitation. *Pediatrics, 77,* 16-28.

Olds, D.L., Robinson, J., O'Brien, R., Luckey, D.W., Pettitt, L.M., Henderson Jr., C.R., … Talmi, A. (2002). Home visiting by paraprofessionals and by nurses: A randomized controlled trial. *Pediatrics, 110*(3), 486-496.

O'Neill, D.K. (2007). The Language Use Inventory for Young Children: A parent-report measure of pragmatic language development for 18- to 47-month-old children. *Journal of Speech, Language, and Hearing Research, 50*, 214-228.

Ortiz-Mantilla, S., Choudhury, N., Leevers, H., Benasich, A.A. (2008). Understanding language and cognitive deficits in very low birth weight children. *Developmental Psychobiology, 50*, 107-126.

Pearson. Bayley Scales of Infant and Toddler Development, Third Edition (Bayley-III). Retrieved from http://www.pearsonassessments.com/HAIWEB/Cultures/enus/Productdetail.htm?Pid=015-8027-23X.

Pearson. Kaufman Assessment Battery for Children, Second Edition (KABC-II). Retrieved from http://www.pearsonclinical.com/psychology/products/100000088/kaufman-assessment-battery-for-children-second-edition-kabc-ii.html#tab-details

Pearson. Mullen Scales of Early Learning. http://www.pearsonclinical.com/childhood/products/100000306/mullen-scales-of-early-learning.html

Pearson. Preschool Language Scales, Fifth Edition (PLS-5). http://www.pearsonclinical.com/language/products/100000233/preschool-language-scales-fifth-edition-pls-5.html

Pearson. Vineland Adaptive Behavior Scales, Second Edition (Vineland-II). http://www.pearsonclinical.com/psychology/products/100000668/vineland-adaptive-behavior-scales-second-edition-vineland-ii-vineland-ii.html

Pearson. Wechsler Preschool and Primary Scale of Intelligence - Fourth Edition (WPPSI-IV). http://www.pearsonclinical.com/psychology/products/100000102/wechsler-preschool-and-primary-scale-of-intelligence--fourth-edition-wppsi-iv.html

Peacock, S., Konrad, S., Watson, E., Nickel, D., & Muhajarine, N. (2013). Effectiveness of home visiting programs on child outcomes: A systematic review. *BMC Public Health, 13*(1), 17.

Pesco, D., & O'Neill, D.K. (2012). Predicting later language outcomes from the Language Use Inventory. *Journal of Speech, Language, and Hearing Research, 55*, 421-434.

Pro-ed. IDA: Infant-Toddler Developmental Assessment. Retrieved from http://www.proedinc.com/customer/productView.aspx?ID=4513

Pro-ed. Early Language Milestone Scale (ELM Scale-2). Retrieved from http://www.proedinc.com/customer/productView.aspx?ID=784

Rescorla, L., & Achenbach, T.M. (2002). Use of the Language Development Survey (LDS) in a national probability sample of children 18 to 35 months old. *Journal of Speech, Language, and Hearing Research, 45*, 733-743.

Rescorla, L., Ratner, N.B., Jusczyk, P., & Jusczyk, A.M. (2005). Concurrent validity of the Language Development Survey: Associations with the MacArthur-Bates Communicative Development Inventories: Words and Sentences. *American Journal of Speech-Language Pathology, 14*, 156-163.

Rescorla, L., & Alley, A. (2001). Validation of the Language Development Survey (LDS): A parent report tool for identifying language delay in toddlers. *Journal of Speech, Language, and Hearing Research, 44*, 434-445.

Riverside Publishing. Battelle Developmental Inventory, Second Edition (BDI-2). Retrieved from http://www.riversidepublishing.com/products/bdi2/

Riverside Publishing. Stanford-Binet Intelligence Scales for Early Childhood (Early SB5). Retrieved from http://riverpub.com/products/earlySB5/

Rydz, D., Shevell, M.I., Majnemer, A., & Oskoui, M. (2005). Developmental screening. *Journal of Child Neurology, 20*(1), 4-21.

Siegel, L.S. (1981). Infant tests as predictors of cognitive and language development at two years. *Child Development, 52,* 545-557.

Scattone, D., Raggio, D. J., & May, W. (2011). Comparison of the Vineland Adaptive Behavior Scales, Second Edition, and the Bayley Scales of Infant and Toddler Development, Third Edition. *Psychological Reports, 109*(2), 626-634.

Schonhaut, L., Armijo, I., Schönstedt, M., Alvarez, J., & Cordero, M. (2013). Validity of the Ages and Stages Questionnaires in term and preterm infants. *Pediatrics, 131*(5), e1468-e1474.

Schwarz, D. F., O'Sullivan, A. L., Guinn, J., Mautone, J. A., Carlson, E. C., Zhao, H., … & Radcliffe, J. (2012). Promoting early intervention referral through a randomized controlled home-visiting program. *Journal of Early Intervention, 34*(1), 20–39

Skarakis-Doyle, E., Campbell, W., & Dempsey, L. (2009). Identification of children with language impairment: investigating the classification accuracy of the Macarthur–Bates Communicative Development Inventories, Level III. *American Journal of Speech-Language Pathology, 18*, 277-288.

Wagner, M., Clayton, S., Gerlach-Downie, S., & McElroy, M. (1999). *An evaluation of the northern California Parents as Teachers demonstration*. Menlo Park, CA: SRI International.

Williams, M.E., Hutchings, J., Bywater, T., Daley, D., & Whitake, C.J. (2013). Schedule of Growing Skills II: Pilot study of an alternative scoring method. *Psychology, 4*(3), 143-152.

Williams, M.E., Sando, L., & Soles, T.G. (2014). Cognitive tests in early childhood: Psychometric and cultural considerations. *Journal of Psychoeducational Assessment, 32*(5) 455-476.

WPS. Cognitive Assessment of Young Children (CAYC). Retrieved from http://www.wpspublish.com/store/p/2701/cognitive-assessment-of-young-children-cayc

WPS. Developmental Profile 3 (DP-3). Retrieved from http://www.wpspublish.com/store/p/2743/developmental-profile-3-dp-3

WPS. Merrill-Palmer-Revised (M-P-R). Retrieved from http://www.wpspublish.com/store/p/2854/merrill-palmer-revised-m-p-r

Yen-Ting Yu, Y.T., Hsieh, W.S., Hsu, C.H., Chen, L.C., Lee, W.T., Chiu, N.C., … Jeng, S.F. (2013). A psychometric study of the Bayley Scales of Infant and Toddler Development – 3rd Edition for term and preterm Taiwanese infants. *Research in Developmental Disabilities, 34*, 3875-3883.

Zimmerman, I.L., & Castilleja, N.F. (2005). The role of a language scale for infant and preschool assessment. *Mental Retardation and Developmental Disabilities Research Reviews, 11*, 238-246.

Appendix A

Child cognitive and language instruments for use in children found in this review

Examiner-administered instruments					
Name	**Type of instrument; Domains assessed**	**Age range**	**Administration time**	**Qualifications required**	**Notes**
Battelle Developmental Inventory, Second Edition (BDI-2)	Examiner-administered (primarily) • Personal-Social • Adaptive • Motor • Communication • Cognitive ability	Birth to 7 years, 11 months	60-90 minutes for entire test	Some training is required for test administrators: general college-level training in measurement and statistical concepts	Recognized as a reliable and valid instrument; shorter screening version available
BRIGANCE Early Childhood Screens III	Examiner-administered • Physical development • Language • Academic/cognitive • Self-help • Social-emotional skills	3 versions: 0-35 months, 3-5 years, 5-6 years	10-15 minutes	Some training required: can be administered by paraprofessionals	Primarily functions as a developmental screen
Cognitive Assessment of Young Children (CAYC)	Examiner-administered • Fine motor coordination and planning • Communication and play • Memory reasoning • Perceptual development • Processing classification and organization • Concept development • Practical knowledge	2 months to 5 years, 11 months	15-30 minutes for entire test	Some training is required: Bachelor's degree recommended	Evidence suggests that validity, reliability, and standardization sample not as highly regarded as BDI-2 or Bayley-III
Denver II Assessment	Examiner-administered • Personal-social • Fine motor • Gross motor-adaptive • Language	Birth to 6 years	20-30 minutes for entire test	2-day training from a master instructor is recommended	Found to have high sensitivity (83%), but an unacceptably low specificity (43%)

Developmental Profile 3 (DP-3)	Examiner-administered • Physical • Adaptive Behavior • Social-Emotional • Cognitive • Communication	Birth to 12 years, 11 months	20-40 minutes for entire test	Some training is required: Bachelor's degree recommended	Appears to be an outdated instrument - limited published evidence, especially recent studies
Early Language Milestone Scale-2 (ELM Scale-2)	Examiner-administered • Speech and language development: • Auditory expressive (subdivided into Content and Intelligibility) • Auditory receptive • Visual	Birth to 36 months	1-10 minutes for entire test	Some training required: professionals having a degree or license to practice in healthcare, childhood specialists	43 items, screens for language delay
Griffiths Mental Development Scales – Revised (GMDS 0-2)	Examiner-administered • Locomotor • Personal-social • Hearing and language • Eye and hand coordination • Performance	0 to 2 years	50 to 60 minutes for entire test	Some training is required: Recommended to have certified training and experience in a relevant discipline, competence in administering psychological tests, and completion of training in Griffiths Mental Development Scales (GMDS) or Autism Diagnostic Observation Schedule (ADOS)	Standardized in the UK in 1997
Kaufman Assessment Battery for Children, Second Edition (KABC-II)	Examiner-administered • Cognitive ability	3-18 years	25-70 minutes, depending on model	High training required: Doctorate degree recommended, with training in administration, scoring, and interpretation of clinical assessments	Age range not applicable for this review
Merrill-Palmer-Revised (M-P-R)	Examiner-administered • Cognitive development • Language/communication • Motor development; • Social–emotional behavior • Self-Help/adaptive behavior	1 month to 6 years, 5 months	45 minutes	Training required: Master's degree recommended, with formal educational training in assessing children, or having a degree or license to practice in healthcare	Validity, reliability, standardization sample not as highly regarded as BDI-2 or Bayley-III

Mullen Scales of Early Learning (MSEL)	Examiner-administered • Gross motor • Visual reception • Fine motor • Expressive language • Receptive language	Birth to 68 months	25-35 minutes for the entire test, when administered to 3 year old children	Training required: Master's degree recommended, with formal educational training in assessing children, or having a degree or license to practice in healthcare	Standardization and normative data outdated - from 1981 and 1989
Peabody Picture Vocabulary Test, Fourth Edition (PPVT-4)	Examiner-administered • Receptive and expressive vocabulary performance	2 years 6 months and up	10-15 minutes	Training is required for test administrators: Master's degree or graduate level training, as well as some familiarity with completing assessments and administering tests, or having a degree or license to practice in healthcare	Age range not applicable for this review
Preschool Language Scale, Fifth Edition (PLS-5)	Examiner-administered • Total language • Auditory comprehension • Expressive communication	Birth to 6 years, 11 months	45-60 minutes for children >1 year	Training is required: Master's degree or graduate level training, as well as some familiarity with completing assessments and administering tests, or having a degree or license to practice in healthcare	Previous edition, PLS-4, widely used and considered a valid and reliable instrument
Schedule of Growing Skills, Second Edition (SGS-II)	Examiner-administered • Passive posture • Active posture • Locomotor • Manipulative • Visual • Hearing and language • Speech and language • Interactive social • Self-care social	Birth to 5 years	20-30 minutes for full test	Some training is required: Professionals can get training sessions from publisher	Standardized in the UK, validated against the Griffiths Mental Development Scales
Stanford–Binet Intelligence Scales for Early Childhood (Early SB5)	Examiner-administered • Intelligence and cognitive abilities: • Fluid reasoning • Knowledge • Quantitative reasoning • Visual-spatial processing • Working memory	2 to 5 years, 11 months	30-50 minutes for full test	Training is required: training in completing, administering, interpreting, and reporting psychological tests	Lacking evidence for use in 2 year old children in terms of validity and reliability - some evidence suggests test is insensitive to small variations in ability at this age

Wechsler Preschool and Primary Scales of Intelligence	Examiner-administered • Verbal reasoning • Concept formation • Sequential processing • Auditory comprehension • Cognitive flexibility • Social judgment • Perceptual organization • Processing speed	2 years, 6 months to 7 years, 3 months	30-40 minutes for ages 2 years, 6 months to 3 years, 11 months	High training required: Doctorate degree recommended, with training in administration, scoring, and interpretation of clinical assessments	Age range not applicable for this review
Examiner-administered & parent-report instruments					
Bayley Scales of Infant and Toddler Development, Third Edition (Bayley-III)	Parent-report & examiner-administered Examiner-administered: • Cognitive • Motor • Language Parent-report: • Interaction • Social-emotional • Adaptive behaviour	1 to 42 months	Up to 90 minutes for entire test for children >13 months	Training is required for test administrators: Master's degree or graduate level training, as well as some familiarity with completing assessments and administering tests, or having a degree or license to practice in healthcare	Recognized as the gold standard test to assess child development; shorter screening version available
Infant-Toddler Developmental Assessment (IDA)	Parent-report & examiner-administered • Gross motor • Fine motor • Relationship to inanimate objects (cognitive) • Language/communication • Self-help • Relationship to persons • Emotions and feeling states (affects) • Coping	Birth to 3 years, 6 months	Dependent on age	Some training is required: Recommended to have training from the Leader's Guide and three videos from the IDA Institute	Aims to identify children who are developmentally at risk; Occurs in 6 stages: Referral & Preinterview, Initial Parent Interview, Health Review, Developmental Observation and Assessment, Integration and Synthesis, Share Findings, Completion, and Report
Parent-report instruments					
Ages and Stages Questionnaires, Third Edition (ASQ-3)	Parent-report • Communication • Gross motor • Fine motor, • Problem solving • Personal-social	1 to 66 months	10-15 minutes for parents to complete 2-3 minutes for professionals to score	Minimal training required to score questionnaires; however, training sessions available with fee from publisher	Primarily functions as a developmental screen; roughly 30 items to screen all 5 domains

Language Development Survey (LDS)	Parent-report • Expressive vocabulary • Word combinations	18 months to 2 years, 11 months	10 minutes for parents to complete	Minimal training required for administrators: administrators typically score the completed forms and generate scores using the instrument materials	Used for language delay screening
Language Use Inventory for Young Children (LUI)	Parent-report • Social pragmatic language • Expressive language	18 months to 3 years, 11 months	20 minutes for parents to complete 5 minutes to score	Minimal training required for administrators to score completed forms	Recently standardized, Canadian instrument
MacArthur-Bates Communicative Development Inventory (CDI)	Parent-report • Language and communication skills	8 to 3 years, 1 month CDI: Words and Sentences for 16-30 months	20-40 minutes for parents to complete 10-15 minutes to score	Minimal training required for administrators: administrators typically score the completed forms and generate scores using the instrument materials	Widely used instrument; many variations/adaptations exist
Parents' Evaluation of Developmental Status (PEDS)	Parent-report • Learning • Development • Behaviour	0 to 8 years	5 minutes for parents to complete 2 minutes to score	Some training required: training available from publisher	Primarily functions as a screening test; Sensitivity =74-80%, specificity = 70%-80%
Vineland Adaptive Behavior Scales, Second Edition (Vineland-II)	Parent-report • Communication • Daily living skills • Socialization Motor skills • Maladaptive behaviour index (optional)	Birth to 90 years	20-60 minutes for full test	Training is required: Master's degree, formal educational training specific to assessing children, or a degree or license to practice in the healthcare	Communication scores found to be significantly higher than corresponding scores from Bayley-III

Appendix B

Child cognitive and language instruments used in home visiting programs that provide services for children at 24 months of age (from Home Visiting Evidence of Effectiveness)*

Program	Study design	Instrument used to assess child cognition or language development	1) Total sample size (group sample sizes if available) 2) Statistically significant difference found between groups? 3) Administered at 24 months?
Child FIRST	RCT (Lowell et al., 2011)	• Infant-Toddler Developmental Assessment (IDA)	1. Total n=117 2. Yes ($p<0.05$) 3. Administered at 6 & 12 months
Early Head Start-Home Visiting	RCT (Love et al., 2001, 2005)	• Bayley Scales of Infant and Toddler Development, Mental Development Index (BSID-MDI) • MacArthur-Bates Communicative Development Index (CDI) • Peabody Picture Vocabulary Test (PPVT)	BSID-MDI: 1. Program n=779, control n=879, 2. No ($p>0.05$) 3. Administered at ~37 months CDI: 1. Total n = 966; 2. No ($p>0.05$) 3. Administered at 24 months PPVT not designed for use at 24 months – not applicable to the BCHCP
Early Start (New Zealand)	RCT (Fergusson et al., 2005)	• Wechsler Preschool and Primary Scale of Intelligence (WPPSI)	WPPSI not designed for use at 24 months – not applicable to the BCHCP
Healthy Families America	RCT (Caldera et al., 2007)	• Ages and Stages Questionnaire (ASQ) • Bayley Scales of Infant and Toddler Development, Mental Development Index (BSID, MDI) • Preschool Language Scale (PLS) - 3 • Stanford Binet Intelligence Scales	ASQ: 1. Total n=513 2. No ($p>0.05$) 3. Administered at 6 months BSID-MDI: 1. Program n=126; control n=123 2. Yes ($p<0.05$) 3. Administered at 24 months PLS-3 1. Total n=513 2. Statistical significance not reported 3. Administered at 36 months
Healthy Steps	RCT (Johnston et al., 2006)	• MacArthur-Bates Communicative Development Index (CDI)	1. Program n=126; control, n=219 2. No ($p>0.05$) 3. Administered at 24 months
MOM Program	RCT (Schwarz et al., 2012)	• Denver Developmental Screening Test-II (DDST-II) • Wechsler Preschool and Primary Scale of Intelligence (WPPSI)-III	DDST-II: 1. Program n=152, control n=150 2. No ($p>0.05$) 3. Administered at 16 months WPPSI-III not designed for use at 24 months – not applicable to the BCHCP

Parent-Child Home Program	RCT (Madden et al., 1984)	• Peabody Picture Vocabulary Test (PPVT) • Stanford Binet Intelligence Scales	Standford Binet Intelligence Scales: 1. Total n=166 2. No (p>0.05) 3. Administered at 4-5 years PPVT not designed for use at 24 months – not applicable to the BCHCP
Parents as Teachers	RCT (Drotar et al., 2009, Wagner et al., 1999)	• Bayley Scales of Infant and Toddler Development, Mental Development Index (BSID, MDI) • Denver Developmental Screening Test (DDST) - II • Developmental Profile (DP) - II • Kaufman Assessment • Peabody Picture Vocabulary Test (PPVT) • Preschool Language Scale (PLS)	BSID-MDI: 1. At 12 months: program n=189, control n=187; at 24 months: program n=166, control n=178 2. No (p>0.05) 3. Administered at 12 & 24 months DP-II: 1. At 12 months: program n=175, control n=140; at 24 months: program n=220, control n=155; at 26 months: program n=210, control n=153 2. No (p>0.05) 3. Administered at 12, 24, & 36 months Kaufman Assessment: 1. Sample size varied for subscales: Program n=141-161, control n=154-170 2. No (p>0.05) 3. Administered at 36 months DDST-II: 1. Total n=206 2. No (p>0.05) 3. Administered at 36 months PLS: 1. Total n=40 2. Yes (p<0.05) 3. Administered at 4-5 years PPVT not designed for use at 24 months – not applicable to the BCHCP
Play and Learning Strategies	RCT (Landry et al., 2008)	• Peabody Picture Vocabulary Test (PPVT) - 3 • Preschool Language Scale (PLS)	PLS-3: 1. Total n=166 2. No (p>0.05) 3. Administered at ~12 months PPVT-3 not designed for use at 24 months – not applicable to the BCHCP

*Note that these programs provide services to children at 24 months old, but these children may be assessed for child cognitive and language development at a later age

Appendix C

Child cognitive and language instruments used in randomized controlled trials (RCT) to evaluate the Nurse-Family Partnership

Authors; RCT location	Instrument used to assess child cognition or language development
Olds et al., 1986; Elmira, USA	• Bayley Scales of Infant and Toddler Development, Mental Development Index (BSID) - II
Kitzman et al., 1997; Memphis, USA	• Bayley Scales of Infant and Toddler Development (BSID) - II
Olds et al., 2002; Denver, USA	• Bayley Scales of Infant and Toddler Development, Mental Development Index (BSID, MDI) - II • Preschool Language Scale (PLS) - 3
Mejdoubi et al., 2011; Netherlands	• Child language assessed at 18 months, but instrument(s) not published
Owen-Jones et al., 2013; UK	• Early Language Milestone Scale (ELM) • Schedule of Growing Skills (SOGS) - II

Appendix D

Glossary of terms

Term	Definition
Coefficients (for reliability and validity)	A numeric value between 0 and 1, which refers to the performance of an instrument's reliability (measured in reliability coefficient, r), or an instrument's validity or correlation with other validated instruments Interpretation of coefficients for the current review (interpretations vary in the research literature): • ≥ 0.90: excellent • 0.80-0.90: good • 0.70-0.80: acceptable • 0.60-0.70: questionable • 0.50-0.60: poor • <0.50: unacceptable
Cognition	Refers to the processes involved in mental abilities, including thinking, problem solving, reasoning, memory, and perception
Concurrent validity	The degree to which the scoring of an instrument is similar to the scoring of an existing instrument that has been established to measure the same domain/variable
Expressive language	The accurate use of language, in areas including grammar, words, and sentences
Inter-rater reliability	The degree to which different examiners/observers are consistent in their scoring and observations while using the same instrument
Language comprehension	The overall ability to understand language in its written and spoken forms, as well as sign language; includes receptive and pragmatic language, accurate syntax and semantics; language comprehension is supported by cognitive ability
Normative data/reference data	The data of the baseline distribution that an instrument's scores are derived from; data is typically from the standardization sample
Pragmatic language	The ability to use language appropriately in social interactions and situations
Predictive validity	The degree to which an instrument can predict future performance in the domain/variable that the instrument is designed to measure
Receptive language	The ability to understand and interpret language that is read or heard
Reliability	The degree at which an instrument can provide consistent results; common forms of reliability are test-retest reliability and inter-rater reliability
Semantics	Refers to the meaning of language, in areas including the use of words, and phrases
Sensitivity	Refers to the proportion of actual cases of a delay/defect/disease that an instrument can correctly detect; with low sensitivity, false-negatives can occur, where the instrument is unable to detect a case that it is designed to detect
Specificity	Refers to the proportion of cases without a delay/defect/disease that an instrument can correctly conclude as absent; with low specificity, false-positives can occur, where the instrument incorrectly concludes that a delay/defect/disease is present when in reality it is absent
Standardization/norming sample	A random sample that is representative of the general population that the instrument is designed to assess
Syntax	Refers to the principles of sentence structure
Test-retest reliability	The degree to which the scores of an instrument are consistent over time
Validity	The degree to which an instrument measures what it is designed to measure; common types include concurrent validity, construct validity, and predictive validity

Appendix E

Components of required resources, administration, and operations used to estimate total budgets for each shortlisted instrument

Component	Estimated value/quantity*
Sample size	500 children
Scientific field interviewer pay rate	Estimated at $25/hour
Examiner training	Examiner-administered instruments: estimate of 1-day training at $1000 with a trained psychologist/psychometrist Parent-reported instruments: Estimate of 1-day internal training to gain familiarity and expertise in using the instrument materials (8 hours at scientific field interviewer rate)
Instrument materials	Variable for the different instruments (see Table 1), possible savings dependent on how materials are bought (separately or in kits; and if scientific field interviewers will be sharing materials) Examiner-administered instruments: for a conservative total budget estimate, assuming that 10 full kits are to be purchased for 10 scientific field interviewers
Practice sessions	Variable for the different instruments Examiner-administered instruments: Estimate of 2 weeks internal training to gain familiarity and expertise in using the instrument materials (at scientific field interviewer rate) Parent-reported instruments: Estimate of 1 week internal training to gain familiarity and expertise in using the instrument materials (at scientific field interviewer rate)
Examiner training follow-up after practice (to refine skills)	Examiner-administered instruments: estimate of 1-day follow-up training at $1000 with a trained psychologist/psychometrist Parent-reported instruments: Estimate of 1-day internal follow-up training (8 hours at scientific field interviewer rate)
Instrument administration time and scoring time	Variable for the different instruments (see Table 1/Appendix A) – note: administration and scoring times may be optimized to shorter intervals with practice and experience with the instruments For a conservative budget, times in Table 1/Appendix A were used to calculate cost, as a product with the scientific field interviewer rate
Examiner correspondence time with BCHCP study team	Variable (not included in total cost estimate)
Examiner mileage costs to conduct home interviews	Variable (not included in total cost estimate)
Contingency	Variable (not included in total cost estimate)

*1 day = 8 hour work day