Sensitivity, Specificity and Predictive Capacity of the Bayley-III in Full Term Children

Miranda Willet

This thesis is submitted in partial fulfilment

of the degree

of Master of Psychology (Clinical)

School of Psychology

University of Adelaide

September 2019

DECLARATION

This thesis contains no material which has been accepted for the award of any other degree or diploma in any University, and, to the best of my knowledge, this report contains no materials previously published except where due reference is made.

I give permission for the digital version of my thesis to be made available on the web, via the University of Adelaide's digital research repository, the Library Search, and also through web search engines, unless permission has been granted by the School to restrict access for a period of time.

[redacted]

Miranda Willet

September 2019

Table of Contents

List of Tables and Figures

Sensitivity, Specificity and Predictive Capacity of the Bayley-III in Full Term Children

Miranda Willet

This literature review is submitted in partial fulfilment

of the degree

of Master of Psychology (Clinical)

School of Psychology

University of Adelaide

September 2019

Word Count: 3816

Abstract

This paper reviews the existing literature on the predictive capacity of the Bayley Scales of Infant Development, 3rd edition (Bayley-III). A number of areas of concern have been identified including the tendency of the Bayley-III to overestimate ability, thereby underestimating delay. Research pertaining to the predictive capacity of the Bayley-III is reviewed in the context of the sensitivity and specificity, relating to delay. Difficulties in the assessment of development are explored and discrepancies between test scores on the Bayley-III and other measures are addressed in this context. Limitations of the current body of research, such as its focus on preterm populations, are presented and recommendations for further research are discussed.

An ability to detect developmental delay in children is essential, in that early intervention can dramatically reduce or prevent a range of difficulties relating to behavioural, social, cognitive and educational domains (Anderson & Burnett, 2016; Spittle, Orton, Anderson, Boyd & Doyle, 2015). However, development is a highly complex and variable process, posing a challenge to clinicians in the diagnosis of delay in areas such as language, behavioural, motor and cognitive development. At present, standardised instruments administered by clinicians are considered the “gold standard” approach for evaluating development in children. One such measure, The Bayley Scale of Infant Development, will be reviewed, specifically in relation to claims that it overestimates development. Furthermore, the act of quantifying early development has proved problematic and the clinical utility of such measures has been questioned, particularly in relation to their predictive capacity for later developmental outcomes, such as intelligence.

Bayley-III

The Bayley Scale of Infant Development (Bayley) is the most commonly employed instrument for the assessment of early development in childhood, with an aim “to identify children with developmental delay and to provide information for intervention planning” (Bayley, 2006, p. 1). Originally published in 1969, the original Bayley included two assessment indices. The Psychomotor Developmental Index, with a focus on gross and fine motor skills and the Mental Development Index, assessing cognitive, language and social skills. In addition to these two direct assessment indices the Behaviour Rating Scale or Infant Behaviour Record was utilised to assess the child’s behaviour throughout the assessment (Bayley, 1969). The second edition was published in 1993 and saw a revision of content, removing items deemed unsatisfactory and including over one-hundred additional items. In addition, normative data was updated, the sets of items were administered according to the child’s age, age range for administration was expanded to 1-42months from 2-30months, and data pertaining to a range of clinical groups was reported in order to enhance clinical utility (Bayley, 1993). However, both the Bayley-II and Bayley lacked clinical utility in that delay on the Mental Development Index scale was unable to be attributed to either cognitive or language delay, as was the case with a differentiation between gross and fine motor delay on the Psychomotor Development Index scale (Moore, Johnson, Haider, Hennessy, & Marlow, 2012).

In 2006, the third edition, Bayley-III was the result of a substantial revision to the structure of the previous editions and significant restandardisation. In restandardising the Bayley-III, 10% of the Bayley-III normative sample of 1,700 infants, was comprised of various clinical populations including Cerebral Palsy, Down Syndrome and Language Impairment, which were not included in the standardisation of the previous editions of the scale (Yi, Sung & Yuk, 2018). Though there is sound justification for this addition, such as increasing the representativeness of a demographic group's characteristics and not truncating the normal distribution, it has been argued that the inclusion of such populations in the normative sample, may be detrimental to the ability of a measure to identify developmental disorders (Pena, Spaulding & Plante, 2006).

The restructure included the formation of five distinct scales; Cognitive; Language; Motor; and caregiver ratings of Adaptive and Social-Emotional behaviour. Of note is the separation of the earlier Mental Developmental Index into distinct Cognitive and Language Indices. Theories and research regarding information processing have formed the basis of the Cognitive Index. As such items measure habituation to visual and auditory stimuli, attention, processing speed and novelty preference (Månsson, Stjernqvist, Serenius, Ådén, & Källén, 2018). Conversely, the Language Index describes the social aspects of communication and is comprised of items that ascertain the ability to attend to and initiate social routines, verbal skills and word comprehension (Månsson et al., 2018).

In addition, the aforementioned issue relating to specificity was addressed by the formation of a subtest for both fine and gross motor tasks, as well as, expressive language, receptive language and cognitive skills, each subtest with its own normative data (Bayley, 2006a). This change allows for a more comprehensive assessment of an individual's relative strengths and weaknesses, thereby facilitating more targeted interventions, increasing clinical utility of the measure (Anderson & Burnett, 2016). A disadvantage of the revision lies in the difficulty experienced by clinicians in a comparison of scores from the Bayley-III to the previous editions and an extended administration time of the revised scale (Anderson & Burnett, 2016). Norms employed for interpretation are largely based on populations and studies conducted in the United States, however data is emerging pertaining to norms standardised in countries such as Germany and the Netherlands (Steenis, Verhoeven, Hessen, & van Baar, 2015).

Overestimation of Bayley-III Test Scores

Despite improvements to the measure, a study conducted by Anderson and colleagues (2010) highlighted concerns regarding the sensitivity of Bayley-III Cognitive, Language and Motor scales; specifically, a tendency of the measure to under-detect developmental delay. Findings of this study, conducted on a population of Australian full-term and preterm two year olds, revealed that, on average, individuals scored .5 standard deviations higher on the Cognitive and Language scales of the Bayley-III and over 1 standard deviation higher on the Motor scale, when compared to the published normative means. Though Anderson and colleagues (2010) acknowledge potential mediating factors such as sampling characteristics or geographic variability, this finding is supported by anecdotal evidence from clinician's use of the Bayley-III in practice (Anderson et al., 2010). Subsequent studies, employing global populations including from the United Kingdom (Moore et al., 2012), Brazil (Silveira, Filipouski, Goldstein, O'Shea, & Procianoy, 2012) and the United States (Vohr et al., 2012), echo this finding, in which significantly higher rates of delay were recorded on the Bayley-II when compared to the Bayley-III.

Moore and colleagues (2012) conducted a similar study to that of Anderson and colleagues, in the United Kingdom, instead administering both the Bayley-II and Bayley-III to the same sample of children born preterm at a median age of 33 months. The sample scored on average 3 points higher on the Cognitive composite scale of the Bayley-III than their MDI score on the Bayley-II, however scores on the Language composite scale were on average 10 points higher than the mean MDI score. In addition, rates of impairment also differed significantly, increasing from 14% on the Bayley-III to 25% on the Bayley-II. Further concerns with the Bayley-III were highlighted in this study by the lower correlation of MDI scores to Cognitive and Language scores at lower score when compared to higher scores, as evidenced by poorer agreement and increasingly greater differences of the averaged cognitive and language scores, as scores fell. Thus, the magnitude of this overestimation looks to fluctuate according to performance. This finding poses significant implications for the use of the Bayley-III as it is children with lower scores who are of most importance in this context as they are most in need of intervention. Limitations of this study lie in the administration of the Bayley-II only items directly after the Bayley-III creating potential attentional and/or test fatigue issues. Further, as there are similarities between the Bayley-II and Bayley-III, there is the potential for learning effects to have impacted upon performance on the Bayley-III.

A smaller study conducted by Silveira and colleagues (2012) on a cohort of sixty children born preterm in Brazil, attempted to address potential attentional or learning effects evident in previous studies. In this study, the Bayley-III and Bayley-II were administered on separate occasions within a two-month time period. Findings of this study mimicked that of previous studies (Moore et al., 2012; Vohr et al., 2012), revealing an 8-9 point difference in mean scores on the Bayley-III Cognitive and Language scales compared to the Bayley-II MDI scale. Despite sound evidence pertaining to the tendency of the Bayley-III to under-report developmental delay, at present all research done in the area has employed a methodology in which the administration of the Bayley-II preceded administration of the Bayley-III. As there is a larger degree of overlap between the two measures, there is still a possibility that a portion of the noted difference between scores could be attributable to learning effects (Anderson & Burnett, 2016).

Vohr and colleagues' (2012) study was conducted in order to compare cognitive outcomes in a cohort of extremely low birth weight children assessed using the Bayley-II (cohort 1) to a cohort assessed using the Bayley-III (cohort 2). Findings revealed that average composite scores were 11 points higher on the cognitive scale, and language composite scores 7 points higher in children assessed in cohort 2 than in cohort 1. In addition, rates of impairment, defined by a score lower than 70, also differed between the two groups, with significantly lower rates in cohort 2. Though demographic differences between cohorts may account for some of these noted differences, further studies, reviewed above, have been conducted in which both the Bayley-II and Bayley-III have been administered to the same cohort, and similar results have been found.

Reinforced by this body of research, it appears there is satisfactory support for the claim that the Bayley-III under-reports developmental delay by overestimating development across its scales (Anderson et al., 2010; Moore et al., 2012; Silveira et al., 2012; Vohr et al., 2012). The magnitude of this overestimation looks to fluctuate according to performance, with evidence pertaining to lower scores producing greater overestimation, posing serious consequences for the identification and provision of suitable interventions (Moore et al., 2012). In addition, though there is evidence of this effect in both children approximately two

(Anderson et al., 2010; Silveira et al., 2012; Vohr et al., 2012) and three years of age (Moore et al., 2012), the degree of overestimation may too vary with age and at present, evidence pertaining to the trajectory of this overestimation not available. The overestimation of scores on the Bayley-III and subsequent underestimation of delay poses serious implications for a clinician's capacity to implement appropriate interventions for those children who require it. Early intervention can dramatically reduce or prevent a range of difficulties relating to behavioural, social, cognitive and educational domains (Anderson & Burnett, 2016; Spittle et al., 2015). As such it is pivotal that delay is being identified in early developmental assessment.

Rates of impairment, according to the Bayley-II, align with the normal distribution of impairment. As such, the assumption that may be drawn from the aforementioned research is that the Bayley-III overestimates development, rather that the contrary explanation that the Bayley-II underestimates development. However, as is evident this research is primarily based on children born preterm therefore evidence pertaining to this trend in full term populations will strengthen the assumption as the true rate of developmental delay in full term children is better known. Children born preterm score significantly lower on the cognitive, receptive language and expressive language domains when assessed at 30months corrected age than their full term peers (Månsson & Stjernqvist, 2014). Furthermore, in Australia in 2017, babies born preterm (<37 weeks) accounted for just 8.7% of live births, posing significant implications for the generalizability of current findings to full term infants (Australian Institute of Health and Welfare, 2019). Anderson and colleagues (2010), attempted to address this limitation in the literature by conducting a study on a cohort of pre-term children with a full term control group in Australia, assessed at twenty-four months of age. Findings of this study revealed significantly higher scores on each cognitive, language and motor composite scales of the Bayley-III than normative means. In addition, rates of impairment (mild to severe delay characterised by a composite score less than 85) are normatively described at 16-17% of a population. This study found rates to be significantly lower, with cognitive, language and motor composite rates of delay at 1%, 4% and 2% respectively (Anderson et al., 2010).

Predicting Development

Predicting later intelligence via assessment of cognitive ability in early childhood and infancy is largely considered ineffective in normal populations (Månsson et al., 2018; McCall & Carriger, 1993). There are numerous explanations for this which have been examined in both clinical and research settings. Construction of measures of cognitive ability in infancy may build upon "extensive and systematic observations of typical development in infants", as such measures usually investigate the attainment of explicit developmental milestones (Månsson et el., 2018, p. 2). Consequently, construction of such measures is based on the assumption of the presence or absence of a particular capability at a particular stage of development. This is the primary characteristic that differentiates tests of development from tests of intelligence. Intelligence as a construct reflects multiple components of intelligence, conceptualized using various mental factors. Many measures of intelligence are based on the Cattell–Horn–Carroll theory (Horn & Cattell, 1966), which separates a general intelligence factor into crystallised and fluid intelligence. Crystallised intelligence is comprised of abilities shaped by experience and includes measures of long-term memory or verbal comprehensions. Conversely, fluid intelligence relies less on prior experience and includes measures of abstract problem solving or logic reasoning (Månsson et al., 2018). Some such constructs are not assessable in infancy as they are yet to develop, however other constructs such as reaction time and novelty preference have proven to be valuable predictors of intelligence in later life (Domsch, Lohaus, & Thomas, 2009; Ellingsen, 2016). Measures of reaction time and novelty processing reflect the underlying cognitive functions of information processing and memory, respectively, associated with an individual's intelligence (Månsson et al., 2018).

Normal development is a complex process and can be conceptualized both in increases in abilities as well as in qualitative progress. As such, the interaction between internal factors, such as genetics, and external factors, including the environment, means that though there is general progression in the attainment of both knowledge and skills, there is a large degree of variation in an infant's developmental progress. It is hypothesized that the influence of genetics on an individual's cognitive ability is strengthened by environmental factors (Tucker-Drob, Briley & Harden, 2013). Infants living in high socioeconomic status environments, such as those with higher maternal educational attainment, are more likely to have greater access to activities that foster positive learning experiences. Conversely, infants are less likely to have such experiences living in low socioeconomic status environments (Månsson et al., 2018). Despite cognitive ability being extremely variable in infancy and young childhood, assuming conditions are not altered dramatically, the intelligence quotient is far more stable. Though a developing individual's capacity to comprehend verbal information, problem solve abstract concepts or reason logically is continuously improving, when compared to their same aged peers, their intelligence quotient usually remains relatively stable, allowing for greater predictability across time (Neisser et al., 1996).

Predictive Capacity of the Bayley-III

As has been noted development is a highly variable and complex process, in addition there is a tendency for developmental outcomes to 'level out' as children age, whereby children identified as delayed at a young age may reach the ability of their peers (Anderson & Burnett, 2016, p. 8). This context, combined with the fact the Bayley Scales have not been designed as a predictor of later functioning mean expectations for the predictive validity of the scales should not be excessive. However, as a common use of the Bayley Scales is in programs for developmentally vulnerable children and it is frequently employed as a primary outcome measure for a range studies including both observational and randomised control trials, the sensitivity and specificity of the measure is important and thus has been explored. In doing so, as well as recognizing potentially overstated scores of development when applying the Bayley-III, one must also acknowledge that, more generally, any measures of development relating to IQ and functioning in older children may also carry measurement error, which may too account for some of the noted differences in ability between assessment scores at the various measurement points (Anderson & Burnett, 2016).

Concerns regarding the predictive capacity of the Bayley Scales for constructs such as general intelligence predate the release of the Bayley-III. Hack and colleagues (2005) conducted a large cohort study on a population of babies born preterm/extremely low birth weight. Assessed at twenty months of age on the Bayley-II and at eight years of age on the Kaufman Assessment Battery's Mental Processing Composite, findings from this study revealed that rates of impairment (MDI score of <70) were significantly lower at age eight, 39% at twenty months corrected age to 16% at aged eight (Hack et al., 2005). In addition, there was much greater variability in scores of children who scored lower, than those who scored >85. These findings were replicated in a more current meta-analysis, conducted to investigate the Bayley Scales' predictive capacity of later functioning in babies born preterm/extremely low birth weight (Luttikhuizen dos Santos, de Kieviet, Königs, van Elburg, & Oosterlaan, 2013). Analysis was comprised predominantly of studies employing the Bayley-I and Bayley-II, only one study included utilising the Bayley-III. Findings from data of included studies revealed a correlation of .61 between mental development index scores and individuals' later cognitive functioning.

There is less available evidence pertaining to the specificity and sensitivity of the Bayley-III relative to an individuals' later functioning. In a study conducted by Bode and colleagues (2014), on a cohort of children born preterm and matched controls, the Bayley-III Cognitive and Language scales were administered at two years corrected age followed by the WPPSI-III at four years of age. Findings revealed a strong correlation between Full-Scale IQ assessed at 4 years and the Bayley-III Cognitive and Language scales, reporting correlations of .81 and .78 respectively. However, this result varied according to gestational age at time of birth, reporting higher correlations in infants with lower gestational ages (Bode, D'Eugenio, Mettelman, & Gross, 2014). Findings of this study reported both high sensitivity and specificity. However, classifications were computed according to the mean and standard deviation of the control group rather than test norms, inconsistent with the intended use of the Bayley-III. Therefore, meaningful evaluation of predictive validity should employ test norms.

Spencer-Smith and colleagues (2015), explored the predictive capacity of the Bayley-III, according to test norms in a cohort of infants born preterm. This study was similar in design, administering the Bayley-III to two year old infants, followed by the DAS-II, a measure of general intelligence, at four years of age. When assessed at two years of age, 11% of the sample were identified as having mild to moderate delay according to the Bayley-III. At four years of age, when reassessed using the DAS-III, the rate of impairment increased to 13% on the Nonverbal Reasoning index, 13% on the Verbal index and 17% on the General Conceptual Ability index. Findings revealed that cognitive and language delay identified by the Bayley-III in infants at two years of age (classified according to test norms: <85 and <70 to identify mild/moderate and moderate delay, respectively and local term-born reference data: <95 and <85, respectively, for the Cognitive scale and <94 and <79, respectively, for the Language scale), was likely to persist when assessed at four years of age (classified according to test-norms: scores < -1 SD relative to the mean were classified as mild/moderate delay or impairment, and scores, < -2 SDs were classified as moderate delay or impairment), reflecting high specificity of the Bayley-III. However, children identified as being impaired at four years of age were not recognized as being delayed by the Bayley-III at two years of age, reflecting poor sensitivity (Spencer-Smith, Spittle, Lee, Doyle, & Anderson, 2015). Thus, the research published to date suggest Bayley-III tends to under-report later cognitive impairment. Though, little is known about the predictive capacity of the Bayley-III, in later functioning of full term infants and in older children, assessed according to test norms.

Agreement Between Test Scores

A study by Månsson, Stjernqvist, Serenius, Ådén, & Källén (2018) on a cohort of full term children aimed to examine the agreement between Bayley-III scores at infancy and Intelligence Quotient (IQ) scores of the WISC-IV at 6.5 years of age. Findings suggested that the strongest association with later IQ was found in the Bayley-III Cognitive Index score, both with regard to Full Scale IQ (FSIQ) and General Ability Index (GAI). This finding supports the significance of cognitive skills in infancy as predictors of later IQ. Furthermore, it may be indicative of the structuring of the revised Cognitive Index in the Bayley-III to measure central information skills such as problem solving skills, processing abilities and speed. However, associations between the WISC-IV FSIQ and the Bayley-III Language Index were lower than anticipated, and lower than those found by Bode and colleagues (2014). The Bayley-III Language Index is designed to be a measure of an infant's capacity to interact with others (Bayley, 2006). As such, it includes social and emotional aspects of communication, such as attending to and initiate social routines, along side functional aspects such as vocabulary or word comprehension. Consequently, the Language Index of the Bayley-III assesses the development of communication capacity more broadly than the Verbal Comprehension Index of the WISC-IV, a measure of verbal acquired knowledge and verbal reasoning. As such, there appears to be a need to assess the constructs of language and cognition separately and specifically.

Conclusion

In summary, this review of the Bayley Scales of Infant Development (Bayley-III) indicates a range of concerns surrounding deficits of the measure. Findings from the aforementioned research suggest that the Bayley-III overestimates infant's ability, thereby underestimating delay. Furthermore, when assessing the predictive capacity of the measure, despite high specificity, the Bayley-III reflects poor sensitivity in evaluating trajectories of children's progress, in the context of delay. Both of these factors, contribute to implications for the identification of delay in children and the implementation of appropriate intervention. Research thus far has primarily focused on the Bayley-III in the context of populations of children born preterm or born at extremely low birth weight, and has been carried out using test norms for comparison. To facilitate a comprehensive understanding of these elements of the Bayley-III, future studies are needed that include data pertaining to a full term cohort according to test norms.

As the association between Language subscale scores on the Bayley-III and WISC-IV appear to be weak (Månsson et al., 2018), it is important to implement measures that separate the cognitive and language subscales of the Bayley-III and investigate their longitudinal trajectories separately and specifically. For this reason, scales that measure these constructs independently should be utilized to explore these trajectories.

References

Australian Institute of Health and Welfare (2019). Australia's mothers and babies 2017 - in brief. Canberra: AIHW.

Anderson, P. J. & Burnett, A. *(2016)* Assessing developmental delay in early childhood - concerns with the Bayley-III scales. *The Clinical Neuropsychologist*, 31(2), 371-381. doi:10.1080/13854046.2016.1216518

Anderson, P. J., De Luca, C. R., Hutchinson, E., Roberts, G., Doyle, L. W., & Victorian Infant Collaborative, G. (2010). Underestimation of developmental delay by the new Bayley-III Scale. *Archives of Pediatrics & Adolescent Medicine, 164(4)*, 352–356. doi:10.1001/archpediatrics.2010.20

Bayley, N. (1993). Bayley Scales of infant development, second edition: Manual. San Antonio, TX: The Psychological Corporation.

Bayley, N. (1969). The Bayley Scales of infant development. San Antonio, TX: The Psychological Corporation.

Bayley, N. (2006). Bayley Scales of infant and toddler development, third edition technical manual. San Antonio, TX: Pearson PsychCorp.

Bode, M. M., D'Eugenio, D. B., Mettelman, B. B., & Gross, S. J. (2014). Predictive validity of the Bayley, Third Edition at 2 years for intelligence quotient at 4 years in preterm infants. *Journal of Developmental and Behavioral Pediatrics, 35(9)*, 570–575. doi:10.1097/DBP.0000000000000110

Domsch, H., Lohaus, A., & Thomas, H. (2009). Prediction of childhood cognitive abilities from a set of early indicators of information processing capabilities. Infant Behavior and Development, 32, 91-102. doi:10.1016/j.infbeh.2008.10.006

Ellingsen, K. M. (2016). Standardized assessment of cognitive development: Instruments and issues. In A. Garro (Ed.), Early childhood assessment in school and clinical child psychology. New York, NY: Springer. doi:10.1007/978-1-4939-6349-2_2

Elliott, C. D. (2007). Differential ability scales. 2nd ed. San Antonio (TX): Harcourt Assessment, 27.

Hack, M., Taylor, H. G., Drotar, D., Schluchter, M., Cartar, L., Wilson-Costello, D., & Morrow, M. (2005). Poor predictive validity of the Bayley Scales of infant development for cognitive function of extremely low birth weight children at school age. *Pediatrics, 116(2)*, 333–341. doi:10.1542/peds.2005-0173

Hohm, E., Jennen-Steinmetz, C., Schmidt, M. H., Laucht, M. (2007). Language development at ten months: Predictive of language outcome and school achievement ten years later? *European Child & Adolescent Psychiatry***,** 16(3), 149-156. doi:10.1007/s00787-006-0567-y

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology*, 57, 253-270. doi:10.1037/h0023816

Johnson, S., Moore, T., & Marlow, N. (2014). Using the Bayley-III to assess neurodevelopmental delay: Which cut-off should be used? Pediatric Research, 75, 670-674. doi:10.1038/pr.2014.10

Lean, R. E., Paul, R. A., Smyser, T. A., Smyser, C. D. & Rogers, C. E. (2018). Social Adversity and Cognitive, Language, and Motor Development of Very Preterm Children from 2 to 5 Years of Age. *The Journal of Pediatrics*, 203, 177-184

Luttikhuizen dos Santos, E. S., de Kieviet, J. F., Königs, M., van Elburg, R. M., & Oosterlaan, J. (2013). Predictive value of the Bayley Scales of infant development on development of very preterm/very low birth weight children: A meta-analysis. *Early Human Development, 89(7)*, 487–496. doi:10.1016/j.earlhumdev.2013.03.008

Månsson, J., Stjernqvist, K., Serenius, F., Ådén, U., & Källén, K. (2018). Agreement Between Bayley-III Measurements and WISC-IV Measurements in Typically Developing Children. *Journal of Psychoeducational Assessment*. doi:10.1177/0734282918781431

Månsson, J., Stjernqvist, K. (2014). Children born extremely preterm show significant lower cognitive, language and motor function levels compared with children born at term, as measured by the Bayley-III at 2.5 years. Acta Pædiatrica, 103(5), 504-511. doi: 10.1111/apa.12585

McCall, R. B., & Carriger, M. S. (1993). A meta-analysis of infant habituation and recognition memory performance as predictors of later IQ. *Child Development*, 64, 57-79. doi:10.2307/1131437

Moore, T., Johnson, S., Haider, S., Hennessy, E., & Marlow, N. (2012). Relationship between test scores using the second and third editions of the Bayley Scales in extremely preterm children. *The Journal of Pediatrics, 160(4)*, 553–558. doi:10.1016/j.jpeds.2011.09.047

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., . . . Urbina, S. (1996). Intelligence: Knowns and unknowns. American Psychologist, 51, 77-101. doi:10.1037/0003- 066X.51.2.77

Peña, E. D., Spaulding, T. J. & Plante, E. (2006). The Composition of Normatibe Groups and Diagnostic Decision Making: Shooting Ourselves in the Foot. American Journal of Speech- Langugage Pathology, 15(3), 247-254.

Semel, E., Wiig, E. H. & Secord, W. A. (2003). Clinical evaluation of language fundamentals, fourth edition(CELF-4). Toronto, Canada: The Psychological Corporation/Harcourt Assessment.

Semel, E., Wiig, E. H. & Secord, W. A. (2004). Clinical evaluation of language fundamentals preschool, second edition (CELF-P2). San Antonio, TX: Harcourt Assessment.

Silveira, R. C., Filipouski, G. R., Goldstein, D. J., O'Shea, T. M., & Procianoy, R. S. (2012). Agreement between Bayley Scales second and third edition assessments of very low-birth-weight infants. *Archives of Pediatrics & Adolescent Medicine, 166(11)*, 1075–1076. doi:10.1001/archpediatrics.2012.732

Spencer-Smith, M. M., Spittle, A. J., Lee, K. J., Doyle, L. W., & Anderson, P. J. (2015). Bayley-III cognitive and language scales in preterm children. *Pediatrics, 135(5)*, 1258–1265. doi:10.1542/peds.2014-3039

Spittle, A., Orton, J., Anderson, P. J., Boyd, R., & Doyle, L. W. (2015). Early developmental intervention programmes provided post hospital discharge to prevent motor and cognitive impairment in preterm infants. *Cochrane Database of Systematic Reviews*, *11*, Art. No. CD005495. doi:10.1002/14651858.CD005495.pub4

Steenis, L. J. P., Verhoeven, M., Hessen, D. J., & van Baar, A. L. (2015). Performance of Dutch Children on the Bayley III: A Comparison Study of US and Dutch Norms. *PLoS ONE*, 10(8). doi:10.1371/journal.pone.0132871

Tucker-Drob, E. M., Briley, D. A., & Harden, K. P. (2013). Genetic and environmental influences on cognition across development and context. *Current Directions in Psychological Science*, 22, 349-355. doi:10.1177/0963721413485087

Vohr, B. R., Stephens, B. E., Higgins, R. D., Bann, C. M., Hintz, S. R., Das, A., ... Fuller, J. (2012). Are outcomes of extremely preterm infants improving? Impact of Bayley assessment on outcomes. *The Journal of Pediatrics, 161(2)*, 222–228. doi:10.1016/j.jpeds.2012.01.057

Wechsler, D. (2011). Wechsler Abbreviated Scale of Intelligence -Second Edition. San Antonio, TX: Pearson.

Woods, P. L., Rieger, I., Wocadlo, C. & Gordon, A. (2014). Predicting the outcome of specific language impairment at five years of age through early developmental assessment in preterm infants. *Early Human Development*, 90(10), 613-619. doi.org/10.1016/j.earlhumdev.2014.07.010

Yi, Y. G., Sung, I. Y. & Yuk, J. S. (2018). Comparison of Second and Third Editions of the Bayley Scales in Children With Suspected Developmental Delay. *Annals of Rehabilitation Medicine*, 42(2), 313-320. doi:10.5535/arm.2018.42.2.313

Sensitivity, Specificity and Predictive Capacity of the Bayley-III in Full Term Children

Miranda Willet

This report is submitted in partial fulfilment

of the degree

of Master of Psychology (Clinical)

For Submission to The Journal of Clinical Child and Adolescent Psychology

School of Psychology

University of Adelaide

September 2019

Word Count: 5025

Abstract

Objective: To investigate trajectories of the Cognitive and Language subscale scores of the Bayley Scales of Infant Development, Third Edition (Bayley-III) and to assess the specificity and sensitivity of the Bayley-III to predict abilities in a longitudinal sample of full term children. *Methods:* Longitudinal data from 726 (predominantly full-term) children collected for a previous trial was utilized for this study. Participants were assessed at 18 months on the Bayley-III, at 4 years on the Differential Ability Scales, Second Edition and the Clinical Evaluation of Language Fundamentals, Preschool 2, and at 7 years on the Wechsler Abbreviated Scales of Intelligence, Second Edition and Clinical Evaluation of Language Fundamentals, Fourth Edition. Linear mixed-effects models were performed and sensitivities, specificities, positive predicted values (PPV) and negative predicted values (NPV) for mild cognitive and language delays between 18m and 4y, 18m and 7y and 4y and 7y were calculated. *Results:* Statistically significant differences were found in the mean Cognitive, Language and Non-verbal scores across follow up ages. Results reflected low sensitivity values, high specificity values, moderate positive predictive values and high negative predictive values across all domains assessed. *Conclusions:* Findings suggest that the Bayley-III overestimates ability in populations of full term children. A large proportion of children that have cognitive or language delay at four and seven years of age, were not detected at 18 month developmental assessments.

Introduction

An ability to detect developmental delay in children is essential, in that early intervention can dramatically reduce or prevent a range of difficulties relating to behavioural, social, cognitive and educational domains (Anderson & Burnett, 2016; Spittle, Orton, Anderson, Boyd & Doyle, 2015). Currently, standardised instruments administered by clinicians are considered the "gold standard" approach for evaluating development in children. However, quantifying development is problematic, being a highly complex and variable process, posing a challenge to clinicians in the diagnosis of delay in areas such as language, behavioural, motor and cognitive development. The clinical utility of such measures has been questioned, particularly in relation to their predictive capacity for later development.

First published in 1969, the Bayley Scale of Infant Development (Bayley) is the most commonly employed instrument for the assessment of early development in childhood, with an aim "to identify children with developmental delay and to provide information for intervention planning". The Bayley included the Psychomotor Developmental Index (PDI), with a focus on gross and fine motor skills and the Mental Development Index (MDI), assessing cognitive, language and social skills (Bayley, 1969). The second edition, published in 1993, removed items deemed unsatisfactory and introduced items (Bayley, 2006, p.1). However, both the Bayley-II and Bayley lacked clinical specificity, as delay on the MDI did not differentiate between cognitive or language delay. Similarly, the PDI did not differentiate between gross and fine motor delay (Moore, Johnson, Haider, Hennessy, & Marlow, 2012).

The third edition, published in 2006, (Bayley-III) substantially restructured and restandardised the previous editions, including five distinct scales: Cognitive; Language; Motor; and caregiver ratings of: Adaptive and Social-Emotional behaviour. Notably, the MDI was split into Cognitive and Language Indices. In restandardising the Bayley-III, 10% of the Bayley-III normative sample of 1,700 infants, was comprised of various clinical populations including Cerebral Palsy, Down Syndrome and Language Impairment, which were not included in the standardisation of the Bayley-II (Yi, Sung & Yuk, 2018). Additionally, the Bayley-III addressed the issue of clinical predictive capacity introducing subtests, each with its own normative data, for: fine and gross motor tasks; expressive language; receptive language; and cognitive skills, (Bayley, 2006a), enabling a more comprehensive assessment of an individual's relative strengths and weaknesses, facilitating more targeted interventions, and increasing clinical utility of the measure (Anderson & Burnett, 2016).

A study conducted by Anderson and colleagues (2010) on a population of Australian full-term and preterm two year olds, highlighted concerns regarding the sensitivity of the Bayley-III Cognitive, Language and Motor scales, identifying a tendency of the scales to under- detect developmental delay. On average, compared to the normative mean, individuals scored .5 standard deviations higher on the Bayley-III Cognitive and Language scales and over 1 standard deviation higher on the Motor scale. Although Anderson and colleagues (2010) acknowledged potential mediating factors, including sampling characteristics or geographic variability, this finding is supported by anecdotal evidence from practicing clinicians using the Bayley-III (Anderson et al., 2010).

Subsequent studies, employing global populations including: the United Kingdom (Moore et al., 2012); Brazil (Silveira et al., 2012); and the United States (Vohr et al., 2012), again report significantly higher delay rates recorded using the Bayley-II compared to the Bayley-III, supporting the claim that the Bayley-III under-reports developmental delay by overestimating development across its scales. Evidence of overestimation in children aged approximately two (Anderson et al., 2010; Silveira et al., 2012; Vohr et al., 2012) and three years old (Moore et al., 2012), suggests that the degree of overestimation may vary with age. The potential overestimation of the Bayley-III scores and subsequent underestimation of delay poses serious implications for a clinician's capacity to implement appropriate interventions. Accordingly, it is pivotal that early developmental delay is correctly assessed.

The Bayley-III was not designed as a predictor of later functioning. However, the sensitivity and specificity of the Bayley-III warrants further exploration because it is commonly used both in programs for developmentally vulnerable children, and as a primary outcome measure for a range studies including observational and randomised control trials. Exploration should recognize that measurement error in may also apply to the IQ and functioning tests in older children, contributing to noted ability differences between assessment scores at the applicable measurement points (Anderson & Burnett, 2016).

Spencer-Smith and colleagues (2015), explored the predictive capacity of the Bayley- III, according to test norms in a cohort of infants born preterm. This study was similar in design, to the Bayley-III: assessing two year olds, followed by the Differential Ability Scales, Second Edition (DAS-II) measure of general intelligence, at four years old. At two years old, 11% of the sample was identified as having mild to moderate delay according to the Bayley- III. At four years old, using the DAS-II General Conceptual Ability index, impairment rates increased to 17%. Findings revealed that cognitive and language delay identified by the Bayley- III in two year olds, was likely to persist when assessed at four years old, reflecting high specificity of the Bayley-III. However, children identified as being impaired at four years old were not recognized as being delayed by the Bayley-III at two years old, reflecting poor sensitivity (Spencer-Smith, Spittle, Lee, Doyle, & Anderson, 2015). Thus, the research published to date suggests the Bayley- III tends to under-identify later cognitive impairment. Though, little is known about the predictive capacity of the Bayley-III in later functioning of full term infants, and when reassessed in older children, according to test norms.

A study by Månsson, Stjernqvist, Serenius, Ådén, & Källén (2018) on a cohort of full term children aimed to examine the agreement between the Bayley-III scores at infancy and IQ scores of the WISC-IV at 6.5 years old. Findings suggested that the strongest association with later IQ was found in the Bayley-III Cognitive Index score, both with regard to FSIQ and GAI. This finding supports the significance of cognitive skills in infancy as predictors of later IQ and may be indicative of restructuring the Bayley-III Cognitive Index to measure central information skills such as problem solving skills, processing abilities and speed.

The associations between the Bayley-III Language Index and WISC-IV FSIQ were lower than anticipated, and lower than those found by Bode and colleagues (2014). The Bayley-III Language Index is designed to be a measure of an infant's capacity to interact with others (Bayley, 2006). As such, it includes social and emotional aspects of communication, such as attending to and initiating social routines, alongside functional aspects such as word comprehension. Consequently, the Bayley-III Language Index assesses the development of communication capacity more broadly than the WISC-IV Verbal Comprehension Index. Accordingly, separate and specific assessment of language and constructs is required.

The weak association between the Bayley-III Language and WISC-IV VCI scores (Månsson et al., 2018), it is important to implement measures that separate the Bayley-III cognitive and language subscales and to investigate their longitudinal trajectories separately and specifically. For this reason, cognitive measures as well as specific assessments of language abilities will be utilized to explore these trajectories. Additionally, a Non-verbal score will be derived from the cognitive measures by removing verbal and language based scales, to further separate language and cognition. Furthermore, research on sensitivity and specificity of the Bayley-III has primarily been carried out on populations born preterm. To facilitate a comprehensive understanding of the sensitivity and specificity of the Bayley-III, future studies are needed that include data pertaining to a full term cohort.

Based on the aforementioned literature, the following hypotheses have been formulated:

- Average scores will be lower on the WASI-II at 7 years and DAS-II at 4 years than on the Bayley-III cognitive scale at 18 months.
- Average scores will be lower on the CELF-4 at 7 years and CELF-P2 at 4 years than of the Bayley-III language scale at 18 months.
- A proportion of children identified as delayed at 4 years and 7 years will not be identified as delayed at 18 months.

Methods

Data for this study was utilised from an existing data set gathered by the DOMInO investigation team between 2009 and 2017. Methodology for the trial, 18 month, 4-year and 7-year follow-ups were previously published (Makrides et al., 2010; Makrides et al., 2014; Gould et al., 2017. Written informed consent was obtained and approval granted by the local institutional ethics review boards. The original trial was a double-blind, multicentre randomized control trial comprising of 2399 women, from 5 Australian hospitals, who were less that 21 weeks gestation with singleton pregnancies recruited between 2005 and 2008. The intervention comprised of Docosahexaenoic acid-rich fish oil capsules or matched vegetable oil capsules from study entry to birth. Follow up of 726 children randomly selected from the original cohort was completed in 2009. There were no significant findings relating to the outcome measures of levels of depressive symptoms in mothers at 6 weeks and 6 months post-partum and cognitive and language development in children assessed by the Bayley-III at 18 months. The trial is being utilized as a cohort study as the trial had negligible effects. Ethics approval was granted to utilise this data by the University of Adelaide HREC on 26th October 2018.

Measures

Bayley-III

At 18 months, the Cognitive and Language Composite Scales of the Bayley Scales of Infant and Toddler Development, Third Edition (Bayley-III) were administered to 726 infants. The cognitive scale evaluates habituation to visual and auditory stimuli, attention, processing speed and novelty preference. This score derived the ‘Cognitive’ score at the 18 month follow up. The language scale is a composite of receptive and expressive communication comprised of items that ascertain the ability to attend to and initiate social routines, verbal skills and word comprehension. This score derived the ‘Language’ score at the 18 month follow up. The motor scale, which evaluates both gross and fine motor functioning, as well as the parental report scales of social-emotional behaviour and adaptive behaviour were assessed as secondary outcome measures. Raw scores for each of the scales are standardized using test norms (mean = 100; SD = 15). The standardized scores were classified, according to test norms, into the categories of mild/moderate delay (<85) and moderate delay (<70).

DAS-II

At 4 years, the core subtests of the Differential Ability Scales, Second Edition (DAS-II) were administered to 631 children. The DAS-II comprises a collection of subtests that assess general reasoning and conceptual abilities, which are used to generate a summary of General Conceptual Ability (GCA) which is similar to an Intelligence Quotient. This score derived the ‘Cognitive’ score at the 4 year follow up. Additional summary indices include the Verbal index, which estimates acquired verbal concepts and knowledge. This score derived the ‘Verbal’ score at the 4 year follow up. The Nonverbal Reasoning index estimates complex nonverbal, inductive reasoning requiring mental processing, this score derived the ‘Non-Verbal’ score at the 4 year follow up. Raw scores for each of the scales are standardized using test norms (mean = 100; SD = 15). The standardized scores were classified, according to test norms, into the categories of mild/moderate delay (<85) and moderate delay (<70).

CELF-P2

At 4 years, the Clinical Evaluation of Language Fundamentals Preschool, second edition (CELF-P2) was administered to 558 children, as a comprehensive developmental language assessment. Three core subtests, Sentence Structure, Word Structure and Expressive Vocabulary were administered. Raw scores for each of the scales are standardized using test norms (mean = 100; SD = 15). A Core Language score is derived from these scores, forming the ‘Language’ score at the 4 year follow up for this study. The standardized scores were classified, according to test norms, into the categories of mild/moderate delay (<85) and moderate delay (<70).

WASI-II

At 7 years, the Wechsler Abbreviated Scales of Intelligence, second edition were administered to 522 children to provide a measure of intelligence. The four subtests; Block Design, Vocabulary, Matrix Reasoning and Similarities were administered. Raw scores for each of the scales are standardized using test norms (mean = 100; SD = 15). A Verbal Comprehension Index score (VCI) is derived from the Vocabulary and Similarities subtest and forms the basis of the ‘Verbal’ score at the 7 year follow up for this study. The Block Design and Matrix Reasoning derive the Perceptual Reasoning Index score (PRI), which translates to the ‘Non-Verbal’ score at the 7 year follow up for this study. All four subtests are combined to form the Full-Scale Intelligence Quotient score (FSIQ-4), which derived the ‘Cognitive’ score at the 7 year follow up for this study. The standardized scores were classified, according to test norms, into the categories of mild/moderate delay (<85) and moderate delay (<70).

CELF-4

At 7 years, the Clinical Evaluation of Language Fundamentals, fourth edition (CELF-4) was administered to 481 children, as a comprehensive developmental language assessment. Four core subtests, Concepts and Following Directions, Word Structure, Recalling Sentences, and Formulating Sentences were administered. Raw scores for each of the scales are standardized using test norms (mean = 100; SD = 15). A Core Language score is derived from these scores, forming the “Language’ score at the 7 year follow up for this study. The standardized scores were classified, according to test norms, into the categories of mild/moderate delay (<85) and moderate delay (<70).

Data Analysis

Analyses were performed using SAS (SAS Institute Inc., Cary, NC, USA), version 9.4. Cognitive, Verbal, Non-verbal and Language scores were each combined across three time periods: 18 months, 4 years and 7 years, from the following separate, age-specific scores. A Cognitive score was derived from the Bayley-III Cognitive composite at 18 months, DAS-II GCA score at 4 years and WASI FSIQ composite at 7 years. A Verbal score was derived from the Bayley-III Language composite at 18 months, DAS-II Verbal score at 4 years and WASI-II Verbal composite at 7 years. A Non-verbal score was derived from the Bayley-III Cognitive composite at 18 months, DAS-II Non-verbal score at 4 years and WASI-II Perceptual composite at 7 years. A Language score was derived from the Bayley-III Language composite at 18 months, CELF-P2 Language score at 4 years and CELF-P4 Standard score at 7 years.

To understand trajectories of scores over time, linear mixed-effects models were performed. These models investigate the association between the outcomes: cognitive, verbal, non-verbal and language scores and the predictor: follow up age, adjusting for repeated measurements over time. An unstructured covariance structure was used to account for measurements taken in each child at three different time points. Cognitive and language mild delays were calculated at each age as ‘Yes’ if score is less than mean minus one standard deviation i.e. <0.85, and 'No' if score is greater than or equal to 0.85. To understand trajectories of delay, sensitivities, specificities, positive predicted values (PPV) and negative predicted values (NPV) for mild cognitive and language delays between 18m and 4y, 18m and 7y and 4y and 7y were calculated.

Results

Table 1. Descriptive statistics for all variables in the analysis.

Characteristic	
Birth weight *g*, mean (SD)	3392 (607)
Gestational age *weeks*, mean (SD)	38.7 (2.2)
Female gender, n (%)	361(49.7)
Race, n (%)	
Caucasian	663 (91.3)
Aboriginal	14 (1.9)
Asian	34 (4.7)
Other	15(2.1)
Maternal Further Education	*484 (66.7)*
English Speaking Home, n (%)	*517(95.9)*
Disorders at 7 Years, n (%)	
Autism Spectrum Disorder	*20 (3.7)*
Attention Deficit Hyperactivity Disorder	*13 (2.4)*
Emotion or Anxiety Disorder	*14 (2.6)*
Speech, Hearing, Auditory or Language Disorder	*60 (11.1)*
Behavioural Disorder	*7 (1.3)*
Learning Disorder	*11 (2.0)*

Table 2. Descriptive statistics of cognitive, verbal, non-verbal and language scores and delays by follow up age.

Variable	*18 months*	*4 years*	*7 years*
Cognitive, mean (SD)	102.0 (11.8)	99.9 (13.2)	97.8 (12.4)
Non-verbal, mean (SD)	102.0 (11.8)	98.4 (13.7)	97.1 (12.5)
Language, mean (SD)	97.4 (14.5)	94.2 (14.5)	93.7 (15.2)
Verbal, mean (SD)	97.4 (14.5)	98.3 (11.5)	98.9 (13.5)
Mild Cognitive delay, n (%)	31 (4.5)	78 (12.4)	69 (13.2)
Mild Language delay, n (%)	118 (17.1)	140 (25.1)	114 (23.7)
Moderate Cognitive delay, n (%)	6(0.8)	17(2.7)	6(1.1)
Moderate Language delay, n (%)	18(2.6)	29(5.2)	29(6.0)

Cognitive – Bayley-III Cognitive composite (18 months), DAS-II GCA score (4 years), WASI Full Scale IQ composite (7 years);
Non-verbal – Bayley-III Cognitive composite (18 months), DAS-II Non-verbal score (4 years), WASI-II Perceptual comp (7 years);
Language – Bayley-III Language composite (18 months), CELF-P2 Language score (4 years), CELF-P4 Standard score (7 years);
Verbal – Bayley-III Language composite (18 months), DAS-II Verbal score (4 years), WASI-II Verbal composite (7 years).

Table 3. Estimates, 95% confidence intervals (CI) and comparison P values of cognitive, verbal, non-verbal and language scores

Outcome	*Comparison*	*Estimate (95% CI)*	*Comparison P value*
Cognitive	18m vs 4y	2.40 (1.36, 3.44)	<.0001
	18m vs 7y	4.88 (3.76, 6.00)	<.0001
	4 vs 7y	2.48 (1.57, 3.39)	<.0001
Non Verbal	18m vs 4y	3.75 (2.60, 4.90)	<.0001
	18m vs 7y	5.34 (4.19, 6.49)	<.0001
	4 vs 7y	1.59 (0.40, 2.78)	0.0088
Language	18m vs 4y	3.61 (2.46, 4.76)	<.0001
	18m vs 7y	4.74 (3.47, 6.01)	<.0001
	4 vs 7y	1.13 (0.03, 2.24)	0.0447
Verbal	18m vs 4y	-0.54 (-1.58, 0.50)	0.3101
	18m vs 7y	-0.79 (-2.06, 0.49)	0.2279
	4 vs 7y	-0.25 (-1.26, 0.76)	0.6295

Cognitive – Bayley-III Cognitive composite (18 months), DAS-II GCA score (4 years), WASI FSIQ comp (7 years);
Non-verbal – Bayley-III Cognitive comp (18 months), DAS-II Non-verbal score (4 years), WASI-II Perceptual comp (7 years);
Language – Bayley-III Language comp (18 months), CELF-P2 Language score (4 years), CELF-P4 Standard score (7 years);
Verbal – Bayley-III Language composite (18 months), DAS-II Verbal score (4 years), WASI-II Verbal comp (7 years).

There is a statistically significant difference in the mean cognitive score across follow up ages, adjusting for repeated measures over time (global P value <0.0001, Table 3). Children at 18 months have a mean cognitive score 4.9 units higher than children at 7 years (estimate=4.9, 95% CI: 3.8, 6.0, comparison P value<0.0001). Further significant post-hoc comparisons can be made between 18 months and 4 years, and between 4 years and 7 years.

There is a statistically significant difference in the mean non-verbal score across follow up ages, adjusting for repeated measures over time (global P value <0.0001, Table 3). Children at 18 months have a mean non-verbal score 5.3 units higher than children at 7 years (estimate=5.3, 95% CI: 4.3, 6.5, comparison P value<0.0001). Further significant post-hoc comparisons can be made between 18 months and 4 years, and between 4 years and 7 years.

There is a statistically significant difference in the mean language score across follow up ages, adjusting for repeated measures over time (global P value <0.0001, Table 3). Children at 18 months have a mean language score 4.7 units higher than children at 7 years (estimate=4.7, 95% CI: 3.5, 6.0, comparison P value<0.0001, Table 3). Further significant post-hoc comparisons can be made between 18 months and 4 years, and between 4 years and 7 years.

There no statistically significant difference in the mean verbal score across follow up ages, adjusted for repeated measures over time (global P value=0.4591, Table 3).

Table 4. Sensitivities, specificities, positive predicted values (PPV) and negative predicted values (NPV) for cognitive and language delays between 18m and 4y, 18m and 7y and 4y and 7y.

	Response variable	*Test variable*	*Test*	*Test value (95% CI)*
18m vs 4y	Cognitive Delay 18m	Cognitive Delay 4y	Sensitivity	0.13 (0.05, 0.20)
	Cognitive Delay 18m	Cognitive Delay 4y	Specificity	0.99 (0.98, 1.00)
	Cognitive Delay 4y	Cognitive Delay 18m	NPV	0.89 (0.86, 0.91)
	Cognitive Delay 4y	Cognitive Delay 18m	PPV	0.56 (0.33, 0.79)
	Language Delay 18m	Language Delay 4y	Sensitivity	0.29 (0.21, 0.36)
	Language Delay 18m	Language Delay 4y	Specificity	0.89 (0.86, 0.92)
	Language Delay 4y	Language Delay 18m	NPV	0.79 (0.75, 0.83)
	Language Delay 4y	Language Delay 18m	PPV	0.46 (0.36, 0.57)
18m vs 7y	Cognitive Delay 18m	Cognitive Delay 7y	Sensitivity	0.13 (0.05, 0.21)
	Cognitive Delay 18m	Cognitive Delay 7y	Specificity	0.98 (0.96, 0.99)
	Cognitive Delay 7y	Cognitive Delay 18m	NPV	0.88 (0.85, 0.91)
	Cognitive Delay 7y	Cognitive Delay 18m	PPV	0.47 (0.25, 0.70)
	Language Delay 18m	Language Delay 7y	Sensitivity	0.32 (0.23, 0.41)
	Language Delay 18m	Language Delay 7y	Specificity	0.89 (0.86, 0.92)
	Language Delay 7y	Language Delay 18m	NPV	0.81 (0.77, 0.85)
	Language Delay 7y	Language Delay 18m	PPV	0.47 (0.36, 0.59)

Cognitive – Bayley-III Cognitive composite (18 months), DAS-II GCA score (4 years), WASI Full Scale IQ composite (7 years); Language – Bayley-III Language composite (18 months), CELF-P2 Language score (4 years), CELF-P4 Standard score (7 years).

The percentage of children with cognitive delay at 18 months who also have cognitive delay at 4 years (sensitivity) is 13% (95% CI: 5%, 20%, Table 4). The percentage of children without cognitive delay at 18 months who also have no cognitive delay at 4 years (specificity) is 99% (95% CI: 98%, 100%, Table 4). The probability that cognitive delay is present at 4 years when there is cognitive delay at 18 months (Positive Predictive Value) is 56% (95% CI: 33%, 79%, Table 4). The probability that cognitive delay is not present at 4 years when there is no cognitive delay at 18 months (Negative Predictive Value) is 89%

(95% CI: 86%, 91%, Table 4).

The percentage of children with cognitive delay at 18 months who also have cognitive delay at 7 years (sensitivity) is 13% (95% Confidence interval (CI): 5%, 21%, Table 4). The percentage of children without cognitive delay at 18 months who also have no cognitive delay at 7 years (specificity) is 98% (95% CI: 96%, 99%, Table 4). The probability that cognitive delay is present at 7 years when there is cognitive delay at 18 months (Positive Predictive Value) is 47% (95% CI: 25%, 70%, Table 4). The probability that cognitive delay is not present at 7 years when there is no cognitive delay at 18 months (Negative Predictive Value) is 88% (95% CI: 85%, 91%, Table 4).

The percentage of children with language delay at 18 months who also have language delay at 4 years (sensitivity) is 29% (95% Confidence interval (CI): 21%, 36%, Table 4). The percentage of children without language delay at 18 months who also have no language delay at 4 years (specificity) is 89% (95% CI: 86%, 92%, Table 4). The probability that language delay is present at 4 years when there is language delay at 18 months (Positive Predictive Value) is 46% (95% CI: 36%, 57%, Table 4). The probability that language delay is not present at 4 years when there is no language delay at 18 months (Negative Predictive Value) is 79% (95% CI: 75%, 83%, Table 4).

The percentage of children with language delay at 18 months who also have language delay at 7 years (sensitivity) is 32% (95% Confidence interval (CI): 23%, 41%, Table 4). The percentage of children without language delay at 18 months who also have no language delay at 7 years (specificity) is 89% (95% CI: 86%, 92%, Table 4). The probability that language delay is present at 7 years when there is language delay at 18 months (Positive Predictive Value) is 47% (95% CI: 36%, 59%, Table 4). The probability that language delay is not present at 7 years when there is no language delay at 18 months (Negative Predictive Value) is 81% (95% CI: 77%, 85%, Table 4).

Discussion

This study aimed to investigate score trajectories of the Cognitive and Language subscales of the Bayley-III and to assess the specificity and sensitivity of the Bayley-III in a longitudinal sample of predominantly full term children. As per our hypotheses a statistically significant difference was found between the mean score on both the cognitive and language scales of the Bayley-III and corresponding scores on the DAS-II and CELF-P4 at 4 years and the WASI-II and CELF-2 at 7 years of age, that derived the Cognitive and Language score comparisons. Participants scored between 2.4 and 3.7 points higher on the Bayley-III at 18 months than at 4 years, and between 4.7 and 5.3 points higher than at 7 years. Moreover, the magnitude of this score difference increased at 7 years compared to 4 years on all three of the aforementioned comparisons. This pattern was echoed when the Cognitive scale of the Bayley-III was compared to Non-verbal indices of the DAS-II and WASI-II to generate a Non-verbal score comparison. This finding supports the notion that the Bayley-III overestimates ability and this degree of overestimation increases with age of follow-up (Anderson et al., 2010; Moore et al., 2012; Silveira et al., 2012; Vohr et al., 2012).

In support of our hypotheses, rates of delay were not stable between 18 months, 4 years and 7 years. Similar findings relating to the sensitivity and specificity of the Bayley-III, to those established in preterm populations, were replicated in this study of predominantly full term infants (Spencer-Smith, 2015). The Cognitive scales of the Bayley-III reflected poor sensitivity, in that children identified as being impaired at four and seven years of age were not recognized as being delayed by the Bayley-III at 18 months of age. However, delay identified by the Bayley-III in infants at 18 months of age, was likely to persist when assessed at four years and seven years of age, reflecting high specificity of the Bayley-III. The Language scales of the Bayley-III echoed this pattern, however reflected higher sensitivity than the Cognitive scales. These findings suggest that early classification of cognitive and language delay is not a reliable predictor of later cognitive functioning.

Findings from analysis of Positive Predictive Values indicate that of those recognized as delayed on the cognitive scales of the Bayley-III at 18 months, 56% and 47% are recognised as delayed, at 4 years and 7 years respectively. This reflects the conclusion that 44% and 53% of those delayed at 4 and 7 years respectively, are not being recognised as delayed at 18 months. Therefore, a large proportion of children that may require developmental intervention would not receive it, should they not be assessed at a later age. As such, poor sensitivity and low Positive Predictive Values of the Bayley-III hold important implications for intervention for delay in young children. These findings support the notion the Bayley-III underestimates delay and emphasise the need for developmental testing to be carried out at an older age, such as at 4 years old, so those children not identified as delayed at 18 months can still be detected, and appropriate intervention applied. Alternatively, clinicians should be looking towards other developmental measures or refining the Bayley-III to eliminate issues of overestimation of ability and poor predictive capacity.

The statistically significant difference found between the mean score on the language scale of the Bayley-III and corresponding scores on the CELF-4 at 4 years and the CELF-2 at 7 years of age, was not found in the Verbal comparison. This comparison was generated through the Language scale of Bayley-III corresponding to the Verbal score of the DAS-II at 4 years and the Verbal composite of the WASI-II at 7 years. This outcome is in support of Månsson and colleagues' findings relating to the agreement between the Language score of the Bayley-III and the Verbal index of the WISC-IV, a measure of IQ similar to the DAS-II and WASI-II. This finding suggests that the Language scale of the Bayley-III may be measuring a different construct to the Verbal indices of the DAS-II, WASI-II and other tests of IQ, strengthening support for the need to investigate language and cognition specifically and separately.

A strength of the current study was the large cohort of full-term children, a population for which little research on the Bayley-III has been conducted. Limitations of the current study include that observations are limited to reference values for 18 months, 4 years and 7 years of age. Therefore, caution should be applied when generalizing findings to other time points. Future research should be undertaken in order to expand understanding on the concepts explored in this study across ages. Additionally, the Bayley-III, DAS-II and WASI-II were all standardized in the United States, as such there may be cultural differences in Australia where the measures were administered, though this has not previously been an issue in prior Australian cohorts (Anderson, 2010).

Conclusions

Patterns of the tendency of the Bayley-III to overestimate ability in populations of children born preterm, persist in full term populations, posing implications for early detection of delay in children born full term. A large proportion of children that have cognitive or language delay at four and seven years of age, are not being detected at 18 month developmental assessments. These findings suggest that early classification of cognitive and language delay is not a reliable predictor of later cognitive functioning. Therefore, developmental assessment should be carried out to at least school age in order to expose all children in need of intervention. It appears as though the constructs captured by the Language scale of the Bayley-III do not align with those captured by the Verbal composites of the IQ test employed in the current study, supporting the need to assess language specifically and separately using comprehensive developmental language assessments.

References

Anderson, P. J. & Burnett, A. *(2016)* Assessing developmental delay in early childhood - concerns with the Bayley-III scales. *The Clinical Neuropsychologist*, 31(2), 371-381. doi:10.1080/13854046.2016.1216518

Anderson, P. J., De Luca, C. R., Hutchinson, E., Roberts, G., Doyle, L. W., & Victorian Infant Collaborative, G. (2010). Underestimation of developmental delay by the new Bayley-III Scale. *Archives of Pediatrics & Adolescent Medicine, 164(4)*, 352–356. doi:10.1001/archpediatrics.2010.20

Bayley, N. (1993). *Bayley Scales of infant development, second edition: Manual*. San Antonio, TX: The Psychological Corporation.

Bayley, N. (1969). The Bayley Scales of infant development. San Antonio, TX: The Psychological Corporation.

Bayley, N. (2006). Bayley Scales of infant and toddler development, third edition technical manual. San Antonio, TX: Pearson PsychCorp.

Elliott, C. D. (2007). Differential ability scales. 2nd ed. San Antonio (TX): Harcourt Assessment, 27.

Gould, J. F., Treyvaud, K., Yelland, L. N., Anderson, P. J., Smithers, L. G., Mcphee, A .J. & Makrides, M. (2017). Seven-Year Follow-up of Children Born to Women in a Randomized Trial of Prenatal DHA Supplementation. *JAMA,* 317(11), 1173–1175. doi:10.1001/jama.2016.21303

Makrides, M., Gibson, R. A., McPhee, A. J., Yelland, L., Quinlivan, J. & Ryan, P. (2010). Effect of DHA Supplementation During Pregnancy on Maternal Depression and Neurodevelopment of Young Children: A Randomized Controlled Trial. *JAMA,* 304(15), 1675–1683. doi:10.1001/jama.2010.1507

Makrides, M., Gould, J. F., Gawlik, N. R., Yelland, L. N., Smithers, L. G., Anderson, P. J. & Gibson, R. A. (2014). Four-Year Follow-up of Children Born to Women in a Randomized Trial of Prenatal DHA Supplementation. *JAMA,* 311(17), 1802–1804. doi:10.1001/jama.2014.2194

Månsson, J., Stjernqvist, K., Serenius, F., Ådén, U., & Källén, K. (2018). Agreement Between Bayley-III Measurements and WISC-IV Measurements in Typically Developing Children. *Journal of Psychoeducational Assessment*. doi:10.1177/0734282918781431

Moore, T., Johnson, S., Haider, S., Hennessy, E., & Marlow, N. (2012). Relationship between test scores using the second and third editions of the Bayley Scales in extremely preterm children. *The Journal of Pediatrics, 160(4)*, 553–558. doi:10.1016/j.jpeds.2011.09.047

Peña, E. D., Spaulding, T. J. & Plante, E. (2006). The Composition of Normatibe Groups and Diagnostic Decision Making: Shooting Ourselves in the Foot. American Journal of Speech- Langugage Pathology, 15(3), 247-254.

Semel, E., Wiig, E. H., & Secord, W. A. (2003). Clinical evaluation of language fundamentals, fourth edition(CELF-4). Toronto, Canada: The Psychological Corporation/Harcourt Assessment.

Semel, E., Wiig, E. H. & Secord, W. A. (2004). Clinical evaluation of language fundamentals preschool, second edition (CELF-P2). San Antonio, TX: Harcourt Assessment.

Silveira, R. C., Filipouski, G. R., Goldstein, D. J., O'Shea, T. M., & Procianoy, R. S. (2012). Agreement between Bayley Scales second and third edition assessments of very low-birth-weight infants. *Archives of Pediatrics & Adolescent Medicine, 166(11)*, 1075–1076. doi:10.1001/archpediatrics.2012.732

Spencer-Smith, M. M., Spittle, A. J., Lee, K. J., Doyle, L. W., & Anderson, P. J. (2015). Bayley-III cognitive and language scales in preterm children. *Pediatrics, 135(5)*, 1258–1265. doi:10.1542/peds.2014-3039

Spittle, A., Orton, J., Anderson, P. J., Boyd, R., & Doyle, L. W. (2015). Early developmental intervention programmes provided post hospital discharge to prevent motor and cognitive impairment in preterm infants. *Cochrane Database of Systematic Reviews*, *11*, Art. No. CD005495. doi:10.1002/14651858.CD005495.pub4

Vohr, B. R., Stephens, B. E., Higgins, R. D., Bann, C. M., Hintz, S. R., Das, A., ... Fuller, J. (2012). Are outcomes of extremely preterm infants improving? Impact of Bayley assessment on outcomes. *The Journal of Pediatrics, 161(2)*, 222–228. doi:10.1016/j.jpeds.2012.01.057

Wechsler, D. (2011). Wechsler Abbreviated Scale of Intelligence -Second Edition. San Antonio, TX: Pearson.

Yi, Y. G., Sung, I. Y. & Yuk, J. S. (2018). Comparison of Second and Third Editions of the Bayley Scales in Children With Suspected Developmental Delay. *Annals of Rehabilitation Medicine*, 42(2), 313-320. doi:10.5535/arm.2018.42.2.313

Appendix A

Instructions for Authors

Journal of Clinical Child and Adolescent Psychology

Preparing Your Paper

Regular Articles, Brief Reports, Future Directions

- Should be written with the following elements in the following order: title page; abstract; main text; references; appendices (as appropriate); table(s) with caption(s) (on individual pages); figures; figure captions (as a list)
- Should contain a structured abstract of 250 words.
- Read making your article more discoverable, including information on choosing a title and search engine optimization.

A Regular Article may not exceed 11,000 words (i.e., 35 pages), including references, footnotes, figures, and tables. Brief Reports include empirical research that is soundly designed, but may be of specialized interest or narrow focus. Brief Reports may not be submitted in part or whole to another journal of general circulation. Brief Reports may not exceed 4,500 words for text and references. These limits do not include the title page, abstract, author note, footnotes, tables, and figures. Manuscripts that exceed these page limits and that are not prepared according to the guidelines in the Manual will be returned to authors without review. Future Directions submissions are written by leading scholars within the field. These articles provide a brief summary of important advances that are needed within a specific research or practice area pertinent to clinical child and adolescent psychology. Future Directions submissions are by invitation only and undergo peer review. All Regular Article and Brief Report submissions must include a title of 15 words or less that identifies the developmental level of the study participants (e.g., children, adolescents, etc.). JCCAP uses an unstructured abstract format. For studies that report randomized clinical trials or meta-analyses, the abstract also must be consistent with the guidelines set forth by CONSORT or MARS, respectively. The Abstract should include up to 250 words, presented in paragraph form. The Abstract should be typed on a separate page (page 2 of the manuscript), and must include each of the following label sections: 1) Objective (i.e., a brief statement of the purpose of the study); 2) Method (i.e., a detailed summary of the participants, N, age, gender, ethnicity, as well as a summary of the study design, measures, and procedures; 3) Results (i.e., a detailed summary of the primary findings that clearly articulate comparison groups (if relevant); 4) Conclusions (i.e., a description of the research and clinical implications of the findings). Avoid abbreviations, diagrams, and reference to the text in the abstract. JCCAP will scrutinize manuscripts for a clear theoretical framework that supports central study hypotheses.

In addition, a clear developmental rationale is required for the selection of participants at a specific age. The Journal is making diligent efforts to insure that there is an appropriately detailed description of the sample, including a) the population from which the sample was drawn; b) the number of participants; c) age, gender, ethnicity, and SES of participants; d) location of sample, including country and community type (rural/urban), e) sample identification/selection; f) how participants were contacted; g) incentives/rewards; h) parent consent/child assent procedures and rates; i) inclusion and exclusion criteria; j) attrition rate. The Discussion section should include a comment regarding the diversity and generality (or lack thereof) of the sample. The Measures section should include details regarding item content and scoring as well as evidence of reliability and validity in similar populations. All manuscripts must include a discussion of the clinical significance of findings, both in terms of statistical reporting and in the discussion of the meaningfulness and clinical relevance of results. Manuscripts should a) report means and standard deviations for all variables, b) report effect sizes for analyses, and c) provide confidence intervals wherever appropriate (e.g., on figures, in tables), particularly for effect sizes on primary study findings. In addition, when reporting the results of interventions, authors should include indicators of clinically significant change. Authors may use one of several approaches that have been recommended for capturing clinical significance, including (but not limited to) the reliable change index (i.e., whether the amount of change displayed by a treated individual is large enough to be meaningful, the extent to which dysfunctional individuals show movement to the functional distribution).

All manuscripts should conform to the criteria listed in Table 1 of the 2008 APA Publications and Communications Board Working Group on Journal Article Reporting Standards (published in American Psychologist). These reporting standards apply to all empirical papers. In addition, JCCAP requires that reports of randomized clinical trials conform to CONSORT reporting standards (http://www.consort-statement.org/index.aspx?o=2965), including the submission of a flow diagram and checklist. Nonrandomized clinical trials must conform to TREND criteria
(see http://www.cdc.gov/trendstatement/docs/AJPH_Mar2004_Trendstatement.pdf) and meta-analyses should conform to MARS standards (see Table 4 in 2008 American Psychologist article).

Style Guidelines
Please refer to these quick style guidelines when preparing your paper, rather than any published articles or a sample copy.
Please use American spelling style consistently throughout your manuscript.
Please use double quotation marks, except where "a quotation is 'within' a quotation". Please note that long quotations should be indented without quotation marks.

Formatting and Templates
Papers may be submitted in Word format. Figures should be saved separately from the text. To assist you in preparing your paper, we provide formatting template(s).
Word templates are available for this journal. Please save the template to your hard drive, ready for use.
If you are not able to use the template via the links (or if you have any other template queries) please contact us here.

References
Please use this reference guide when preparing your paper.

Taylor & Francis Editing Services
To help you improve your manuscript and prepare it for submission, Taylor & Francis provides a range of editing services. Choose from options such as English Language Editing, which will ensure that your article is free of spelling and grammar errors, Translation, and Artwork Preparation. For more information, including pricing, visit this website.

Checklist: What to Include

1. **Author details.** All authors of a manuscript should include their full name and affiliation on the cover page of the manuscript. Where available, please also include ORCiDs and social media handles (Facebook, Twitter or LinkedIn). One author will need to be identified as the corresponding author, with their email address normally displayed in the article PDF (depending on the journal) and the online article. Authors' affiliations are the affiliations where the research was conducted. If any of the named co-authors moves affiliation during the peer-review process, the new affiliation can be given as a footnote. Please note that no changes to affiliation can be made after your paper is accepted. Read more on authorship.
2. You can opt to include a **video abstract** with your article. Find out how these can help your work reach a wider audience, and what to think about when filming.
3. **Funding details.** Please supply all details required by your funding and grant-awarding bodies as follows:
   For single agency grants
   This work was supported by the [Funding Agency] under Grant [number xxxx].
   For multiple agency grants
   This work was supported by the [Funding Agency <] under Grant [number xxxx]; [Funding Agency >] under Grant [number xxxx]; and [Funding Agency &] under Grant [number xxxx].
4. **Disclosure statement.** This is to acknowledge any financial interest or benefit that has arisen from the direct applications of your research. Further guidance on what is a conflict of interest and how to disclose it.
5. **Data availability statement.** If there is a data set associated with the paper, please provide information about where the data supporting the results or analyses presented in the paper can be found. Where applicable, this should include the hyperlink, DOI or other persistent identifier associated with the data set(s). Templates are also available to support authors.
6. **Data deposition.** If you choose to share or make the data underlying the study open, please deposit your data in a recognized data repository prior to or at the time of submission. You will be asked to provide the DOI, pre-reserved DOI, or other persistent identifier for the data set.
7. **Supplemental online material.** Supplemental material can be a video, dataset, fileset, sound file or anything which supports (and is pertinent to) your paper. We publish supplemental material online via Figshare. Find out more about supplemental material and how to submit it with your article.
8. **Figures.** Figures should be high quality (1200 dpi for line art, 600 dpi for grayscale and 300 dpi for color, at the correct size). Figures should be supplied in one of our preferred file formats: EPS, PDF, PS, JPEG, TIFF, or Microsoft Word (DOC or DOCX) files are acceptable for figures that have been drawn in Word. For information relating to other file types, please consult our Submission of electronic artwork document.
9. **Tables.** Tables should present new information rather than duplicating what is in the text. Readers should be able to interpret the table without reference to the text. Please supply editable files.
10. **Equations.** If you are submitting your manuscript as a Word document, please ensure that equations are editable. More information about mathematical symbols and equations.
11. **Units.** Please use SI units (non-italicized).

Using Third-Party Material in your Paper

You must obtain the necessary permission to reuse third-party material in your article. The use of short extracts of text and some other types of material is usually permitted, on a limited basis, for the purposes of criticism and review without securing formal permission. If you wish to include any material in your paper for which you do not hold copyright, and which is not covered by this informal agreement, you will need to obtain written permission from the copyright owner prior to submission. More information on requesting permission to reproduce work(s) under copyright.

Submitting Your Paper

This journal uses ScholarOne Manuscripts to manage the peer-review process. If you haven't submitted a paper to this journal before, you will need to create an account in ScholarOne. Please read the guidelines above and then submit your paper in the relevant Author Center, where you will find user guides and a helpdesk.

Please note that *Journal of Clinical Child & Adolescent Psychology* uses Crossref™ to screen papers for unoriginal material. By submitting your paper to *Journal of Clinical Child & Adolescent Psychology* you are agreeing to originality checks during the peer-review and production processes.

On acceptance, we recommend that you keep a copy of your Accepted Manuscript. Find out more about sharing your work.

Data Sharing Policy

This journal applies the Taylor & Francis Basic Data Sharing Policy. Authors are encouraged to share or make open the data supporting the results or analyses presented in their paper where this does not violate the protection of human subjects or other valid privacy or security concerns.

Authors are encouraged to deposit the dataset(s) in a recognized data repository that can mint a persistent digital identifier, preferably a digital object identifier (DOI) and recognizes a long-term preservation plan. If you are uncertain about where to deposit your data, please see this information regarding repositories.

Authors are further encouraged to cite any data sets referenced in the article and provide a Data Availability Statement.

At the point of submission, you will be asked if there is a data set associated with the paper. If you reply yes, you will be asked to provide the DOI, pre-registered DOI, hyperlink, or other persistent identifier associated with the data set(s). If you have selected to provide a pre-registered DOI, please be prepared to share the reviewer URL associated with your data deposit, upon request by reviewers.

Where one or multiple data sets are associated with a manuscript, these are not formally peer reviewed as a part of the journal submission process. It is the author's responsibility to ensure the soundness of data. Any errors in the data rest solely with the producers of the data set(s).

Publication Charges

There are no submission fees, publication fees or page charges for this journal.

Color figures will be reproduced in color in your online article free of charge. If it is necessary for the figures to be reproduced in color in the print version, a charge will apply. Charges for color figures in print are $400 per figure (£300; $500 Australian Dollars; €350). For more than 4 color figures, figures 5 and above will be charged at $75 per figure (£50; $100 Australian Dollars; €65). Depending on your location, these charges may be subject to local taxes.

Copyright Options

Copyright allows you to protect your original material, and stop others from using your work without your permission. Taylor & Francis offers a number of different license and reuse options, including Creative Commons licenses when publishing open access. Read more on publishing agreements.

Complying with Funding Agencies
We will deposit all National Institutes of Health or Wellcome Trust-funded papers into PubMedCentral on behalf of authors, meeting the requirements of their respective open access policies. If this applies to you, please tell our production team when you receive your article proofs, so we can do this for you. Check funders' open access policy mandates here. Find out more about sharing your work.

Open Access
This journal gives authors the option to publish open access via our Open Select publishing program, making it free to access online immediately on publication. Many funders mandate publishing your research open access; you can check open access funder policies and mandates here.
Taylor & Francis Open Select gives you, your institution or funder the option of paying an article publishing charge (APC) to make an article open access. Please contact openaccess@tandf.co.uk if you would like to find out more, or go to our Author Services website.
For more information on license options, embargo periods and APCs for this journal please go here.

My Authored Works
On publication, you will be able to view, download and check your article's metrics (downloads, citations and Altmetric data) via My Authored Works on Taylor & Francis Online. This is where you can access every article you have published with us, as well as your free eprints link, so you can quickly and easily share your work with friends and colleagues.
We are committed to promoting and increasing the visibility of your article. Here are some tips and ideas on how you can work with us to promote your research.

Article Reprints
You will be sent a link to order article reprints via your account in our production system. For enquiries about reprints, please contact Taylor & Francis at reprints@taylorandfrancis.com. You can also order print copies of the journal issue in which your article appears.

Queries
Should you have any queries, please visit our Author Services website or contact us here.