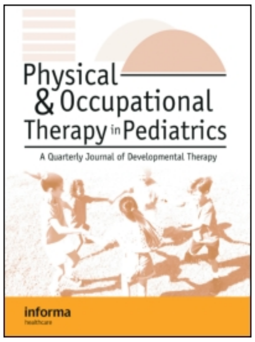
Physical & Occupational Therapy in Pediatrics
A Quarterly Journal of Developmental Therapy
informa

Taylor & Francis
Taylor & Francis Group

Physical & Occupational Therapy In Pediatrics

ISSN: 0194-2638 (Print) 1541-3144 (Online) Journal homepage: http://www.tandfonline.com/loi/ipop20

Reliability of Neurobehavioral Assessments from Birth to Term Equivalent Age in Preterm and Term Born Infants

Abbey L. Eeles, Joy E. Olsen, Jennifer M. Walsh, Emma K. McInnes, Charlotte M.L. Molesworth, Jeanie L.Y. Cheong, Lex W. Doyle & Alicia J. Spittle

To cite this article: Abbey L. Eeles, Joy E. Olsen, Jennifer M. Walsh, Emma K. McInnes, Charlotte M.L. Molesworth, Jeanie L.Y. Cheong, Lex W. Doyle & Alicia J. Spittle (2016): Reliability of Neurobehavioral Assessments from Birth to Term Equivalent Age in Preterm and Term Born Infants, Physical & Occupational Therapy In Pediatrics, DOI: 10.3109/01942638.2015.1135845

To link to this article: http://dx.doi.org/10.3109/01942638.2015.1135845

Published online: 22 Mar 2016.

Physical & Occupational Therapy in Pediatrics, 00(00):1–12, 2016
© 2016 Taylor & Francis Group, LLC
ISSN: 1541-3144 print / 1541-3144 online
DOI: 10.3109/01942638.2015.1135845

Reliability of Neurobehavioral Assessments from Birth to Term Equivalent Age in Preterm and Term Born Infants

Abbey L. Eeles, Joy E. Olsen, Jennifer M. Walsh, Emma K. McInnes, Charlotte M.L. Molesworth, Jeanie L.Y. Cheong, Lex W. Doyle, & Alicia J. Spittle

Victorian Infant Brain Study, Murdoch Childrens Research Institute, RCH, Parkville, Victoria, Australia

ABSTRACT. Neurobehavioral assessments provide insight into the functional integrity of the developing brain and help guide early intervention for preterm (<37 weeks' gestation) infants. In the context of shorter hospital stays, clinicians often need to assess preterm infants prior to term equivalent age. Few neurobehavioral assessments used in the preterm period have established interrater reliability. *Aim:* To evaluate the interrater reliability of the Hammersmith Neonatal Neurological Examination (HNNE) and the NICU Network Neurobehavioral Scale (NNNS), when used both preterm and at term (>36 weeks). *Methods:* Thirty-five preterm infants and 11 term controls were recruited. Five assessors double-scored the HNNE and NNNS administered either preterm or at term. A one-way random effects, absolute, single-measures interclass correlation coefficient (ICC) was calculated to determine interrater reliability. *Results:* Interrater reliability for the HNNE was excellent (ICC > 0.74) for optimality scores, and good (ICC 0.60–0.74) to excellent for subtotal scores, except for 'Tone Patterns' (ICC 0.54). On the NNNS, interrater reliability was predominantly excellent for all items. Interrater agreement was generally excellent at both time points. *Conclusions:* Overall, the HNNE and NNNS neurobehavioral assessments demonstrated mostly excellent interrater reliability when used prior to term and at term.

KEYWORDS. Intraclass correlation coefficient, inter-rater reliability, neurobehavioral assessment, preterm infant

INTRODUCTION AND PURPOSE

The increasing numbers of survivors born preterm or low birth weight are at high-risk of long-term neurodevelopmental problems including cognitive, motor, and behavioral impairments. Up to 50% of those born very preterm are reported to require additional assistance at school age, and there is emerging evidence of the

Address Correspondence to: Dr Abbey L. Eeles, Victorian Infant Brain Study, Murdoch Childrens Research Institute, Level 4, RCH, Parkville, Victoria 3052, Australia (E-mail: abbey.eeles@mcri.edu.au).

(Received 19 April 2015; accepted 18 December 2015)

developmental difficulties associated with late preterm birth (Cheong and Doyle, 2012). Early intervention programs for preterm infants have demonstrated that remediation and rehabilitation improve some developmental outcomes (Kaaresen et al., 2008; Spittle et al., 2007), highlighting the need to identify impairments and provide intervention as early as possible in order to prevent dysfunction or ameliorate adverse neurodevelopmental outcomes. Neurobehavioral assessment of the preterm infant is a window into the functional integrity of the developing brain and can help guide our earliest interventions with this vulnerable population. In the clinical setting, standardized neurobehavioral assessments are commonly used at, or post, term equivalent age, when preterm infants are more robust and stable than prior to term. However, with the growing demand for neonatal intensive care services within tertiary hospitals, preterm infants are often transferred at earlier postmenstrual ages (PMA) to lower level facilities that may not have personnel experienced in neurobehavioral assessment, or are discharged home prior to term equivalent age, before a neurobehavioral assessment can be used to detect early signs of impairment and early intervention referrals can be initiated. Whilst the majority of neurobehavioral assessments identified as appropriate for use in preterm infants have a published lower age limit of around 32 weeks' gestational age (GA; Noble and Boyd, 2012), few have established reliability for use in preterm infants prior to term equivalent age.

Reliability is the process of determining that a test or measure is measuring something in a reproducible and consistent fashion (Law, 2004; Spittle et al., 2008). Interrater reliability (also referred to as interobserver agreement) is the degree of agreement between more than one rater when assessing the same individual at the same time (Kimberlin and Winterstein, 2008).

The presentation of neurobehavior in preterm infants at term equivalent age is reported to vary greatly when compared with their term born peers (Mercuri et al., 2003; Brown et al., 2006a). This variability, and the observational nature of neurobehavioral assessment often require the examiner to identify subtle, qualitative differences in infant behavior. Therefore, in order to have confidence in assessment findings, and to identify changes in neurobehavior that truly exist and are not inflated by measurement error, it is important that these neurobehavioral assessments have established interrater reliability. There are several neurobehavioral assessments recommended for use in newborn infants that vary in purpose (e.g. neurological, behavioral) and clinical utility (e.g. training required, time to administer and score; Brown and Spittle, 2014). Two commonly used neurobehavioral assessments are the Hammersmith Neonatal Neurological Examination (HNNE; Dubowitz et al., 1999) and the NICU Network Neurobehavioral Scale (NNNS; Lester et al., 2004a). Both the HNNE and NNNS were designed for use in preterm and term, high risk infants. Whilst the HNNE incorporates a behavioral component, it is has a strong neurological focus compared to the NNNS, which also evaluates neurological integrity but is behaviorally based. The HNNE is reported to have excellent clinical utility, and is recommended as the assessment of choice if clinicians require a shorter discriminative and predictive assessment (Noble and Boyd, 2012). Several reviews report the HNNE to have clinical utility as it is relatively quick to administer and score, no formal training is required (Wusthoff, 2013; El-Dib, 2011). In contrast, the NNNS is a more comprehensive assessment, taking longer to perform and score and requires formal training and accreditation, which may limit its use in clinical settings, however, it is recommended for use in research or for experienced clinicians (Noble and Boyd, 2012). Dubowitz et al. (1999) describe one of the main applications of the HNNE as serial assessment of preterm infants to document changes in neurological development. The HNNE has good sensitivity (88%) for identifying infants with significant MRI abnormalities, however, its specificity to truly detect infants without significant MRI abnormalities is poor (46%; Woodward et al., 2004a). The HNNE has been used in several studies of very preterm infants as a neurodevelopmental assessment at term (Ricci et al., 2008; Brown et al., 2006b; Woodward et al., 2004b), however, whilst some studies report a procedure was undertaken to ensure agreement between assessors when administering the HNNE, these studies fail to document an interrater reliability value (Dubowitz et al., 1998a; McGready et al., 2000). The NNNS has been shown to be predictive of behavior, motor, and cognitive outcomes (El-Dib et al., 2012; Sucharew et al., 2012; Stephens et al., 2010). Whilst the NNNS is recommended for use from 34 weeks to 48 weeks' corrected age, there are no studies evaluating the interrater reliability of the HNNE for use in preterm infants prior to term equivalent age. Similarly, there have been no studies to date reporting the interrater reliability values of the NNNS for use at term equivalent age and/or prior to term (Fink et al., 2012). In a study aimed to develop norms for a clinically healthy population of full term infants, interrater reliability, whilst not documented in the study's results, was said to be established using criteria recommended for other neurobehavior assessments. Interrater reliability of the NNNS in this study was established based on assessors having no more than two 2-point disagreements on items with 9-point scales and complete agreement on items with 5-point scales. The study did not describe the criteria used with the multiple items on the NNNS that have a scale more than 9 points (maximum 12-point scale) or less than 5 (minimum 4-point scale), or categorical items (present or absent). Furthermore, it is argued that percentages of agreement are an inadequate measure of interrater reliability as they do not correct for agreements that would be expected by chance and thus overestimate the level of agreement (Hallgren, 2012). Accordingly, the purpose of this study is to evaluate the interrater reliability of two neurobehavioral assessments: the Hammersmith Neonatal Neurological Examination (HNNE; Dubowitz et al., 1998b) and the NICU Network Neurobehavioral Scale (NNNS; Lester et al., 2004a), and to examine whether PMA at assessment affects interrater reliability on two neurobehavioral assessments at preterm ($<$37 weeks' PMA) and term equivalent time points ($>$37 weeks' PMA).

METHODS

This reliability study was conducted as part of two longitudinal observational studies documenting the neurobehavioral development of infants born preterm.

Participants

Infants born before 37 weeks' GA and a convenience sample of healthy term controls were recruited from the neonatal intensive care unit (NICU), special care nursery and maternity wards at the Royal Women's Hospital and Frances Perry

House, in Melbourne, Australia, between January 2010 and December 2013. Infants were excluded if they had congenital abnormalities known to affect neurodevelopment or if they were identified as having a poor chance of survival, as assessed by the infants treating medical team. Ethics approval from the human research ethics committees of the hospitals was obtained prior to the commencement of the study and informed consent was obtained from the parents of participants.

Assessments

The HNNE (Dubowitz et al., 1998b) and NNNS (Lester et al., 2004a) were used to assess neurobehavior in preterm infants across varying PMAs, ranging from 30 weeks to term equivalent age (mean: 38; SD: 4.2) and in term control infants between 40–43 weeks' PMA (mean: 41.45; SD: 1). All examiners were experienced in the handling of preterm infants and administering the HNNE, and were certified in the NNNS. All infants were assessed according to the standardized procedures for the two assessments and as previously published in the longitudinal observational study protocol (Spittle et al., 2014). The NNNS was administered first to all infants as it specifies a preferred order of item administration whereas the HNNE has a more flexible approach. The HNNE has multiple items that overlap with the NNNS and thus to avoid unnecessary handling, these HNNE items were scored based on the infants' performance during the NNNS. Once the NNNS was completed, the unique items from the HNNE were delivered to complete the scoring. Assessments administered during the preterm period were timed with the infants daily care procedures, to minimize additional handling. Due to the poor self-regulation abilities and sensitivity to handling of infants at younger gestations, assessments were paced according to the infant's cues and items were omitted (e.g. moro reflex) if they were deemed inappropriate for the infants current state and regulation.

Hammersmith Neonatal Neurological Examination (HNNE; Dubowitz et al., 1998b)

The HNNE is primarily a neurological exam developed for term and preterm infants. It consists of 34 individual items with six subtotals including tone, tone patterns, reflexes, spontaneous movements, abnormal neurological signs, and behavior. It provides an overall "optimality score" which has been validated in healthy term ($n = 224$; Dubowitz et al., 1998b) and preterm ($n = 380$; Ricci et al., 2008) infants.

The NICU Network Neurobehavioral Scale (NNNS; Lester et al., 2004a)

The NNNS provides an in-depth assessment of neurobehavior and is performed on medically stable infants. It assesses the neurological integrity, behavioral functioning, and responses to stress in high-risk infants using 45 items and 13 summary scores including habituation, attention, arousal, regulation, handling, quality of movement, excitability, lethargy, nonoptimal reflexes, asymmetrical reflexes, hypertonicity, hypotonicity, and stress compared with norms for healthy term infants ($n = 344$; Tronick et al., 2004; Fink et al., 2012). The habituation scale was excluded from this analysis as the infants were not consistently in an appropriate state (sleep) to administer the scale at the start of the assessment.

Scoring Procedure

In order to score the HNNE and NNNS for the current interrater reliability study, one clinician administered the assessment whilst the other observed. This process minimized the handling of the infants; however, for items that required manual assessment to score (e.g. muscle tone), the observing clinician repeated the item immediately after the administering clinician. Once the assessment was completed, both clinicians independently scored the assessment. Optimality scores for the HNNE were calculated using the published optimality scores according to PMA at assessment. As there are currently no optimality scores published for preterm infants (<37 weeks' GA), optimality scores for infants assessed before 37 weeks' PMA in the current study were calculated using the published optimality scores for the lowest age band (37–38 weeks' PMA). Summary scores for the NNNS were calculated using the published summary scoring calculations appendix (Lester et al., 2004b).

Statistical Analysis

The appropriate statistics for examining interrater and intrarater reliability is dependent on a study's design but can include Intra Class Correlation (ICC), Lin's Concordance Correlation Coefficient (CCC) or Cohen's Kappa (κ). The ICC is one of the most commonly used statistics for assessing interrater reliability for ordinal, interval and ratio variables and is appropriate for the current study as it accommodates ratings made by a random selection of multiple coders (Hallgren, 2012). With regards to ICC, higher values indicate greater interrater reliability, with an ICC estimate of 1 indicating perfect agreement and 0 indicating only random agreement. Qualitative ratings used to describe the degree of interrater reliability have been established and include ICC values between 0.75 and 1 representing excellent agreement; values between 0.60 and 0.74 classified as good agreement; values between 0.40 and 0.59 considered fair; and values less than 0.40 rated as having poor agreement (Cicchetti, 1994). Interrater reliability was assessed using STATA version 13 and a one-way random effects, absolute, single-measures interclass correlation coefficient calculated to assess the degree that examiners provided agreement in their ratings of the HNNE optimality scores and NNNS summary scores. The use of a one-way model for the ICC was selected to prevent the ICC from accounting for systematic deviations due to specific coders. Furthermore, a single measures variant of ICC was chosen as it is recommended in studies where a subset of subjects is coded by multiple raters and their reliability is used to generalize to subjects rated by one coder (Hallgren, 2012). Lastly, in order to generalize our results to a larger population of therapists administering the neurobehavioral assessments, a random effects variant of ICC was used. Due to binomial data and imbalanced marginals on one of the NNNS summary scores, Hypertonicity, it was not appropriate to use ICC and thus a proportion of total agreement was used to measure inter-rater agreement on this summary score.

We assumed that interobserver agreement would be excellent (ICC $\geq$0.75) on most occasions. With two observers and a desire to limit the width of the 95% confidence interval (CI) for the ICC to be no greater than 0.3 for an ICC of 0.75, we calculated that at least 34 infants would need to be studied. For the smallest subgroup size of 16, the width of the CI would increase to 0.44 for an ICC of 0.75. We enrolled slightly more infants to allow for more variability in the ICCs and sample sizes for the different subtests. ICCs were estimated for the total sample, and also separately for each PMA at neurobehavioral assessment group; term and <37 weeks. Permutation tests were used to assess the difference between PMA group ICCs; the null hypothesis being that ICCs were exchangeable between the two groups (Pesarin and Salmoso, 2010). Ten thousand permutations were used to generate a sampling distribution of ICC differences used to obtain a p-value indicative of the probability that the permuted difference in ICC was at least as great as the observed difference in ICC, assuming no true difference in ICC. At an α level of 5%, a p-value less than 0.05 indicated that the ICC difference was unlikely to be due to chance alone, given no true difference in ICC between the two age groups. In the case of Hypertonicity, a permutation test was performed using the proportion of total agreement.

RESULTS

There were five different assessors administering the HNNE and NNNS in the current study. Forty-six infants were recruited for this study; 35 infants were born preterm (before 37 weeks PMA) and 11 infants were healthy term controls.

Hammersmith Neonatal Neurological Examination (HNNE)

Reliability data for use of the HNNE were available for 38 infants. Twenty-seven infants were born preterm and 11 infants were healthy term controls. Sixteen infants were assessed before 37 weeks' PMA with a mean gestational age of 34.33 (2.02) weeks' at the time of assessment. Twenty-two infants were assessed at term equivalent age (>37 weeks' PMA); 11 preterm and 11 healthy term controls, with a mean gestational age of 41.63 (1.51) at the time of assessment. Infants assessed at preterm age were more likely to be from a multiple birth ($p = <0.01$). There were no significant differences between age of assessment (term vs. preterm) and sex, standardized birth weight, social risk, or medical complications (Table 1).

TABLE 1. Characteristics of Infants Assessed Using the HNNE

	Combined Cohort $n = 38$	Term Assessment (>37 W GA) $n = 22$	Preterm Assessment (<37 W GA) $n = 16$
GA mean (SD)	32 (5)	34 (6)	30 (3)
Birthweight (grams) mean (SD)	1879 (1196)	2291 (1384)	1312 (479)
Male *n* (%)	13 (34)	8 (36)	5 (31)
Multiple *n* (%)	14 (37)	4 (18)	10 (62)
CLD*n* (%)	6 (16)	4 (18)	2 (12)
Suspected / definite NEC *n* (%)	2 (9.5) $n = 21$	2 (20) $n = 10$	0 (0) $n = 11$
IVH *n* (%)	4 (19) $n = 21$	1 (10) $n = 10$	3 (27) $n = 11$
Higher social risk n (%)	14 (39)**	9 (43)*	5 (33)*
Age at assessment mean (SD)	38.56 (4)	41.63 (1.5)	34.33 (2)
Born preterm (<37 weeks' GA) *n* (%)	27 (71.1)	11 (50)	16 (100)

*1 missing value; **2 missing values.

TABLE 2. Individual Intraclass Correlation Coefficients and 95% Confidence Intervals Across the HNNE Subscales and Optimality Scores When Used Before and at Term Equivalent Age

HNNE Subscales	Combined Cohort $n = 38$	Term Assessment (>37 W GA) $n = 22$	Preterm Assessment (<37 W GA) $n = 16$	Permutation Test p-value
Tone	0.88 (0.76 to 0.94) $n = 32$	0.88 (0.73 to 0.95) $n = 20$	0.79 (0.44 to 0.93) $n = 12$	0.482
Tone Patterns	0.54 (0.25 to 0.74) $n = 34$	0.55 (0.16 to 0.79) $n = 20$	0.53 (0.04 to 0.82) $n = 14$	0.855
Reflexes	0.85 (0.69 to 0.93) $n = 26$	0.83 (0.61 to 0.93) $n = 19$	0.93 (0.68 to 0.99) $n = 7$	0.672
Spontaneous Movements	0.82 (0.69 to 0.91) $n = 37$	0.78 (0.53 to 0.90) $n = 21$	0.73 (0.40 to 0.90) $n = 16$	0.693
Abnormal Neurological Signs	0.93 (0.87 to 0.96) $n = 37$	1 $n = 22$	0.86 (0.64 to 0.95) $n = 15$	0.158
Behavior	0.88 (0.77 to 0.94) $n = 33$	0.69 (0.39 to 0.86) $n = 21$	0.95 (0.85 to 0.99) $n = 12$	0.101
Optimality Score	0.94 (0.85 to 0.97) $n = 22$	0.90 (0.76 to 0.96) $n = 18$	0.99 (0.91 to 1) $n = 4$	0.243

Results of the interrater reliability of the HNNE are summarized in Table 2. Overall, ICC values on the optimality score and five out of six subtotal scores on the HNNE indicated excellent agreement (0.82 to 0.94). The subtotal score for "Tone Patterns" yielded fair agreement. When interrater reliability was further evaluated by age at assessment (term vs. preterm), ICC values for four out of six subtotal scores at term equivalent age and the "Optimality Score" had excellent agreement, whilst the "Behavior" subtotal score demonstrated good agreement (0.69) and the "Tone Patterns" subtotal remained fair. Agreement of HNNE subtotal scores when administered at preterm age was excellent on four out of six subtotal scores, including "Behavior" and the "Optimality Score." "Spontaneous Movements" showed good agreement and "Tone Patterns" remained fair at preterm age. The number of infants with complete data differed across the HNNE subscales as an optimality score cannot be generated if there are missing items. The order of administration may have contributed to the missing items on the HNNE as it was administered after the NNNS and infants may have been considered too fatigued for the assessor to continue the handling procedures.

The NICU Network Neurobehavioral Scale (NNNS)

Forty-six infants were assessed using the NNNS. Thirty-five infants were born preterm (<37 weeks) and 11 infants were healthy term controls. At the time of assessment, 18 infants were still preterm (<37 weeks) and 28 were term equivalent age (>37 weeks). Infants assessed at preterm age were more likely to be a multiple birth ($p < 0.01$). There were no other significant differences between infant characteristics and age of assessment (term vs. preterm). Characteristics of infants assessed using the NNNS are summarized in Table 3. All infants had complete data for the NNNS summary scores as missing items are accommodated in the scoring procedure.

TABLE 3. Characteristics of Infants Assessed Using the NNNS

	Combined cohort $n = 46$	Term Assessment (>37 weeks) $n = 28$	Preterm Assessment (<37 weeks) $n = 18$
GA mean (SD)	33 (5)	34 (5)	31 (3)
Birthweight (grams) mean (SD)	1929 (1121)	2238 (91258)	1448 (626)
Male gender n (%)	17 (37)	11 (39)	6 (33)
Multiple n (%)	15 (33)	5 (18)	10 (56)
CLD n (%)	6 (13)	4 (14)	2 (11)
Suspected/definite NEC n (%)	2 (9) $n = 22$	2 (18) $n = 11$	0 (0) $n = 11$
IVH n (%)	4 (18) $n = 22$	1 (9) $n = 11$	2(27) $n = 11$
Higher social risk	16 (36)**	10 (37)*	6 (35)*
Age at assessment mean (SD)	38.84 (4)	41.69 (1.5)	34.40 (2)
Born preterm (<37 weeks' GA n (%)	35 (76.1)	17 (60.7)	18 (100)

*1 missing value; **2 missing values.

Individual ICC and confidence intervals across the NNNS subscales when administered pre and post term equivalent age are summarized in Table 4. With the exception of "Asymmetrical Reflexes," which had good agreement, the ICC values on the combined cohort of infants assessed using the NNNS demonstrated excellent agreement. Scores for infants assessed at term equivalent age had excellent agreement on 10 out of the 12 summary scores, with the exception of "Nonoptimal Reflexes" identified as having good agreement and "Asymmetrical Reflexes" as fair. Similarly, summary scores for preterm infants indicated excellent agreement, with the exception of "Attention" with an ICC of 0.72 demonstrating good agreement.

Overall, there was little evidence that reliability was affected by PMA at assessment. Using permutation tests, there was no evidence that ICCs differed between the two age groups for the HNNE subscales (Table 2: permutation test p-value) nor for most subscales of the NNNs (Table 4: permutation test p-value); however, there was moderate evidence of a difference in ICC, with greater ICC values in the preterm group, for both the Arousal subscale ($p = 0.020$) and the Nonoptimal reflexes subscale ($p = 0.012$).

DISCUSSION

The current study evaluated the interrater reliability of two neurobehavioral assessments when administered both in the preterm period and at term equivalent age. Overall, both the HNNE and NNNS had good interrater reliability, with similar findings prior to term and at term equivalent ages for both preterm and term born infants. The resulting ICCs for the HNNE optimality score at both time points were in the excellent range. Similarly, the six subtotal scores were also in the good to excellent range, with the exception of "Tone Patterns," which demonstrated fair agreement. The reduced interrater reliability on the "Tone Patterns" subtotal score may have been influenced by the observing clinician's reduced handling of the infant and therefore their scoring of tone patterns being predominantly reliant on visual observation. The administering clinician may have scored the infants tone patterns differently due to longer time spent handling the infant. While the HNNE

TABLE 4. Individual Intraclass Correlation Coefficients and 95% Confidence Intervals Across the NNNS Subscales When Used at Varying PMA

	Combined cohort $n = 46$	Term assessment (>37 W GA) $n = 28$	Preterm assessment (<37 W GA) $n = 18$	Permutation test p-value
Attention	0.85 (0.72 to 0.92)	0.91 (0.79 to 0.96)	0.72 (0.33 to 0.90)	0.391
Arousal	0.95 (0.90 to 0.97)	0.87 (0.74 to 0.94)	0.96 (0.89 to 0.98)	0.020
Regulation	0.97 (0.94 to 0.98)	0.97 (0.94 to 0.99)	0.91 (0.76 to 0.97)	0.569
Handling	0.87 (0.76 to 0.93)	0.82 (0.64 to 0.92)	0.95 (0.86 to 0.98)	0.303
Quality of Movement	0.92 (0.86 to 0.95)	0.90 (0.79 to 0.95)	0.95 (0.87 to 0.98)	0.266
Excitability	0.94 (0.89 to 0.97)	0.94 (0.88 to 0.97)	0.87 (0.70 to 0.95)	0.077
Lethargy	0.91 (0.84 to 0.95)	0.88 (0.77 to 0.94)	0.82 (0.58 to 0.93)	0.599
Nonoptimal Reflexes	0.88 (0.79 to 0.93)	0.73 (0.50 to 0.86)	0.97 (0.92 to 0.99)	0.012
Asymmetrical Reflexes	0.67 (0.47 to 0.80)	0.55 (0.23 to 0.76)	0.81 (0.56 to 0.92)	0.267
Hypertonicity*	$p = 0.96$ (0.90 to 1.00)	$p = 0.96$ (0.90 to 1.00)	$p = 0.94$ (0.84 to 1.00)	0.624
Hypotonicity	0.93 (0.87 to 0.96)	0.82 (0.64 to 0.91)	0.93 (0.84 to 0.97)	0.144
Stress	0.86 (0.75 to 0.92)	0.79 (0.59 to 0.90)	0.91 (0.79 to 0.97)	0.261

NB: Numbers differ across subscales due to scores not generated for that subscale.

*p = proportion of total agreement with 95% confidence interval reported for binomial variable with high levels of agreement between ratings.

in the current study was conducted by experienced assessors, these findings provide support for the HNNE authors' assertion that the HNNE can be reliably learnt from the assessment manual.

The ICC for the NNNS when used on infants before 37 weeks' PMA and post 37 weeks' PMA was predominantly in the excellent range, with only a few exceptions. The generally high ICC values on the HNNE optimality scores and NNNS summary scores indicate that clinicians had a high degree of agreement and that both neurobehavioral assessments were scored similarly between clinicians. Furthermore, the high ICC values suggests that a minimal amount of measurement error was introduced by the independent clinicians scoring the neurobehavioral assessments and therefore the statistical power for subsequent analysis using the HNNE optimality scores and NNNS summary scores in the current study will not be substantially reduced. Understanding the interrater reliability is essential in clinical and research settings, as neurobehavioral assessments may be administered by more than one clinician and thus it is important to ensure consistency in assessment findings from which outcomes, therapeutic interventions and referral processes for different infants and families may be based. While the current study did not address the predictive validity of the HNNE and NNNS for later developmental outcome, it provides important evidence for the interrater reliability of these assessments for clinicians who are experienced in handling preterm infants and are certified in the NNNS. Future research is required to investigate the relationship between early neurobehavior and long-term follow up, and thus, the extent to which neurobehavioral assessments prior to term equivalent age may assist clinicians in the earlier detection of neurobehavioral abnormalities, facilitating the earlier application of intervention services to ameliorate longer-term impairments.

CONCLUSIONS

This is the first study to our knowledge to adequately report the interrater reliability of the HNNE and NNNS neurobehavioral assessments. Overall, both assessments demonstrated mostly excellent interrater reliability when used both prior to term and at term equivalent age.

Declaration of Interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the article.

ABOUT THE AUTHORS

Dr **Abbey L. Eeles,** PhD, is a postdoctoral research officer with the Victorian Infant Brain Studies at Murdoch Childrens Research Institute, Melbourne and Senior Occupational Therapist in Newborn Services at Southern Health, Melbourne. Dr **Joy E. Olsen,** PhD, is a postdoctoral research officer with the Victorian Infant Brain Studies at Murdoch Childrens Research Institute, Melbourne and Senior Occupational Therapist in Neonatal Services at the Royal Women's Hospital, Melbourne. Dr **Jennifer M. Walsh,** MD, is a Neonatal Paediatrician in Neonatal Services at the Royal Women's Hospital, Melbourne and the Paediatric, Infant, Perinatal, Emergency Retrieval service at the Royal Children's Hospital, Melbourne. Ms **Emma K. McInnes,** RN, MPH, is a research coordinator with the Victorian Infant Brain Studies at Murdoch Childrens Research Institute, Melbourne. Ms **Charlotte M.L. Molesworth,** MBiostat, is a biostatistician with the Clinical Epidemiology and Biostatistics Unit, Murdoch Childrens Research Institute, Melbourne. Professor **Jeanie L.Y. Cheong,** MD, is a Neonatal Paediatrician in Neonatal Services at the Royal Women's Hospital, Melbourne and Principal Research Fellow at the Murdoch Childrens Research Institute, Melbourne. Professor **Lex W. Doyle,** MD, is the Associate Director of Research at the Royal Women's Hospital, Melbourne and leader of the National Health and Medical Research Committee Centre for Research Excellence in Newborn Medicine. **Dr Alicia J. Spittle,** PhD, is a Principal Research Fellow in the Department of Physiotherapy at The University of Melbourne. She is the leader of the motor team in the Victorian Infant Brain Studies at Murdoch Childrens Research Institute and Senior Physiotherapist in Neonatal Services at the Royal Women's Hospital, Melbourne.

REFERENCES

Brown, N., & Spittle, A. (2014). Neurobehavioral evaluation in the preterm and term infant. *Current Pediatric Reviews*, *10*, 65–72.

Brown, N. C., Doyle, L. W., Bear, M. J., & Inder, T. E. (2006a). Alterations in neurobehavior at term reflect differing perinatal exposures in very preterm infants. *Pediatrics*, *118*, 2461–2471.

Brown, N. C., Doyle, L. W., Bear, M. J., & Inder, T. E. (2006b). Alterations in neurobehavior at term reflect differing perinatal exposures in very preterm infants. *Pediatrics*, *118*, 2461–2471.

Cheong, J. L. Y., & Doyle, L. W. (2012). Increasing rates of prematurity and epidemiology of late preterm birth. *Journal of Paediatrics and Child Health*, *48*, 784–788.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284–290.

Dubowitz, L., Dubowitz, V., & Mercuri, E. (1999). *The neurological assessment of the preterm and full term infant*. London: McKeith Press.

Dubowitz, L., Mercuri, E., & Dubowitz, V. (1998a). An optimality score for the neurologic examination of the term newborn. *The Journal of Pediatrics*, *133*, 406–416.

Dubowitz, L., Mercuri, E., & Dubowitz, V. (1998b). An optimality score for the neurologic examination of the term newborn. *J Pediatr*, *133*, 406–416.

El-Dib, M. (2011). Neurodevelopmental assessment of the newborn: An opportunity for prediction of outcome. *Brain Dev*, *33*, 95–105.

El-Dib, M., Massaro, A. N., Glass, P., & Aly, H. (2012). Neurobehavioral assessment as a predictor of neurodevelopmental outcome in preterm infants. *Journal Of Perinatology: Official Journal Of The California Perinatal Association*, *32*, 299–303.

Fink, N. S., Tronick, E., Olson, K., & Lester, B. (2012). Healthy newborns' neurobehavior: Norms and relations to medical and demographic factors. *Journal of Pediatrics*, *161*, 1073–1079.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*, 23–34.

Kaaresen, P. I., Ronning, J. A., Tunby, J., Nordhov, S. M., Ulvund, S. E., & Dahl, L. B. (2008). A randomized controlled trial of an early intervention program in low birth weight children: outcome at 2 years. *Early Human Development*, *84*, 201–209.

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy*, *65*, 2276–2284.

Law, M. (2004). *Outcome measures rating form.* Ontario, Canada: CanChild Centre for Disability Research.

Lester, B. M., Tronick, E. Z., & Brazelton, T. B. (2004a). The Neonatal Intensive Care Unit Network Neurobehavioral Scale procedures. *Pediatrics*, *113*, 641–667.

Lester, B. M., Tronick, E. Z., & Brazelton, T. B. (2004b). The neonatal intensive care unit network neurobehavioral scale procedures; Appendix 3: Summary score calculations. *Pediatrics*, *113*, 695–699.

Mcgready, R., Simpson, J., Panyavudhikrai, S., & Al, E. (2000). Neonatal neurological testing in resource-poor settings. *Annals of Tropical Paediatrics*, *20*, 323–336.

Mercuri, E., Guzzetta, A., Laroche, S., Ricci, D., Vanhaastert, I., Simpson, A., ... Dubowitz, L. (2003). Neurologic examination of preterm infants at term age: Comparison with term infants. *The Journal of Pediatrics*, *142*, 647–655.

Noble, Y., & Boyd, R. (2012). Neonatal assessments for the preterm infant up to 4 months corrected age: a systematic review. *Developmental Medicine & Child Neurology*, *54*, 129–139.

Pesarin, F., & Salmoso, L. (2010). *Permutation Tests for Complex Data: Theory, Application and Software*. Hoboken, N.J: Wiley.

Ricci, D., Romeo, D. M. M., Haataja, L., van Haastert, I. C., Cesarini, L., Maunu, J., ... Mercuri, E. (2008). Neurological examination of preterm infants at term equivalent age. *Early Human Development*, *84*, 751–761.

Spittle, A. J., Doyle, L. W., & Boyd, R. N. (2008). A systematic review of the clinimetric properties of neuromotor assessments for preterm infants during the first year of life. *Developmental Medicine and Child Neurology*, *50*, 254–266.

Spittle, A. J., Orton, J., Doyle, L. W., & Boyd, R. (2007). Early developmental intervention programs post hospital discharge to prevent motor and cognitive impairments in preterm infants. *Cochrane Database Syst Rev*, 2.

Spittle, A. J., Thompson, D. K., Brown, N. C., Treyvaud, K., Cheong, J. L. Y., Lee, K. J., Anderson, P. J. (2014). Neurobehaviour between birth and 40 weeks gestation in infants born <30 weeks gestation and parental psychological wellbeing: predictors of brain development and child outcomes. *BMC Pediatrics*, *14*(111), 1–13.

Stephens, B. E., Liu, J., Lester, B., Lagasse, L., Shankaran, S., Bada, H., Higgins, R. (2010). Neurobehavioral assessment predicts motor outcome in preterm infants. *J Pediatr*, *156*, 366–371.

Sucharew, H., Khoury, J. C., Xu, Y., Succop, P., & Yolton, K. (2012). NICU Network Neurobehavioral Scale profiles predict developmental outcomes in a low-risk sample. *Paediatr Perinat Epidemiol*, *26*, 344–352.

Tronick, E. Z., Olson, K., Rosenberg, R., Bohne, L., Lu, J., & Lester, B. M. (2004). Normative neurobehavioral performance of healthy infants on the Neonatal Intensive Care Unit Network Neurobehavioral Scale. *Pediatrics*, *113*, 676–678.

Woodward, L. J., Mogridge, N., Wells, S. W., & Inder, T. E. (2004a). Can neurobehavioral examination predict the presence of cerebral injury in the very low birth weight infant? *Journal of Developmental & Behavioral Pediatrics*, *25*, 326–334.

Woodward, L. J., Mogridge, N., Wells, S. W., & Inder, T. E. (2004b). Can neurobehavioral examination predict the presence of cerebral injury in the very low birth weight infant? *Developmental and Behavioral Pediatrics*, *25*, 326–334.

Wusthoff, C. (2013). How to use: the neonatal neurological examination. *Archives of Disease in Childhood, Education and Practice Edition*, *98*, 148–153.