Critical Review: Exploring the inter-rater reliability of two assessment tools used to identify feeding problems in neonates

Gillian Lucas
M.Cl.Sc (SLP) Candidate
University of Western Ontario: School of Communication Sciences and Disorders

This critical review examines the inter-rater reliability of two clinical feeding assessments of infant oral sensorimotor function, the Neonatal Oral-Motor Scale (NOMAS) and the Preterm Infant Breastfeeding Behaviour Scale (PIBBS). Study designs include four diagnostic test studies and one systematic review. Tests of inter-rater reliability for the PIBBS resulted in acceptable agreement between observers but not between observers and mothers in one study, and good or excellent agreement between observers and excellent agreement between observer and mothers in a later investigation. NOMAS reliability studies revealed both moderate to high inter-rater agreement as well as moderate to substantial inter-observer values. Further research is required to continue to validate the psychometric soundness of these instruments before any definitive conclusions can be made regarding their standard use in identification of infant feeding problems and evaluation of neonatal treatment approaches.

Introduction

There has been a dramatic increase in the prevalence of feeding difficulties in the neonatal intensive care unit as a result of the survival of greater numbers of preterm, medically fragile, and chronically ill infants (Palmer, Crawley, & Blanco, 1993). Research suggests that feeding difficulties can result in a wide array of persisting and devastating consequences to the preterm infant. These may include a negative impact on the mother-infant dyad (Hill & Johnson, 2007), failure to thrive, and inadequate nutritional intake, which can lead to behavioural, developmental, and cognitive impairments (Leonard & Kendall, 1997).

Direct assessment of feeding competency (e.g. manometry to assess muscle pressure and movements associated with swallowing), can be invasive and have a damaging effect on the ill newborn, and often requires intricate measuring and analyzing instruments to generate the data (da Costa & van der Schans, 2008). Therefore, a clinical observational assessment which accurately identifies neonates at risk for feeding problems is critical and can provide a framework for early intervention (Leonard & Kendall, 1997).

Although published clinical guides have been developed to aid in the assessment process (Subramaniam, 2001), many currently available tools have not been standardized (Arvedson, 2008), their psychometric properties have not been extensively critiqued and compared (Howe, Lin, Fu, Su, & Hsieh, 2008), and no single assessment is utilized consistently throughout the United States, Canada, or the rest of the world (Rogers & Arvedson, 2005).

The power of an assessment tool is reliant to a great extent by it scientific soundness. That is to say, an effective tool should display scientific integrity in three basic psychometric properties: reliability, validity and responsiveness (Howe, Sheu, Hsieh, & Hsieh, 2007). Reliability refers to the extent which an assessment tool is free from random error (Aaronson, Alonso, Burnam, Lohr, Patrick, Perrin, & Stein, 2002). Evidence on reliability and validity is important in making decisions concerning the potential of an assessment measure for use in clinical settings and in future research studies (da Costa & van der Schans, 2008). One method for examining reliability is determining a tool's inter-rater agreement at one point in time (i.e., the extent to which two different observers obtain a similar outcome when using the same instrument to measure a concept).

It is imperative that clinicians and researchers be cognizant of the psychometric properties of neonatal feeding assessment tools to ensure that they are able to provide early identification of feeding problems, as well as provide reliable monitoring to determine the success of interventions (Howe et al., 2007).

Thus, the purpose of this review is to examine the inter-rater reliability of two feeding assessment tools commonly used with neonates, the Neonatal Oral-Motor Scale (NOMAS) and the Preterm Infant Breastfeeding Behaviour Scale (PIBBS).

Objectives

The primary objective of this paper is to critically evaluate existing literature regarding the inter-rater reliability of the NOMAS and PIBBS; two feeding

assessment tools used in the neonatal population. The secondary objective is to propose evidence-based research and clinical recommendations regarding the use of these tools in pediatric practice.

Methods

Search Strategy

Computerized databases, including CINAHL, SCOPUS, PubMed, and ProQuest Dissertations and Theses were searched using the following search strategy: ((infant feeding) AND (assessment) OR (Neonatal Oral Motor Assessment Scale) OR (Neonatal Oral Motor Assessment Scale) AND (reliability) OR (Preterm Infant Breastfeeding Behaviour Scale) OR (Preterm Infant Breastfeeding Behaviour Scale) AND (reliability) OR (pediatric dysphagia) AND (assessment)).

Selection Criteria

Studies selected for inclusion in this critical review paper were required to specifically investigate the inter-rater reliability of the PIBBS and NOMAS and/or to report on the further development of either of the tools. Treatment studies which utilized either the NOMAS or PIBBS as an outcome measure for determining intervention efficacy were excluded.

Data Collection

Results of the literature search yielded the following types of articles congruent with the aforementioned selection criteria: diagnostic test study (4) and systematic review (1).

Results

In an attempt to categorize the oral-motor patterns that characterize poor feeding in preterm infants, an observational assessment tool was created to describe jaw and tongue function during nonnutritive and nutritive sucking. The Neonatal Oral-Motor Assessment Scale (NOMAS) separates 13 characteristics of jaw movement and 13 characteristics of tongue movement into categories of normal, disorganized, and dysfunctional. The scale was administered to 40 infants by Palmer, Crawley, and Blanco (1993) to establish inter-rater reliability, alter the scale as necessary based on the reliability outcomes, and attempt more detailed description of the patterns of disorganized and dysfunctional sucking in neonates. The percentage of agreement was calculated on each of the 26 items, which were scored by three independent observers. Inter-rater agreement of all 26 items ranged from 63-100%. Ten items were found to be 90% to 100% reliable and seven items were 80-90% reliable. However, given the lack of methodological rigor in this study, this research provides suggestive evidence of the inter-rater reliability of the NOMAS.

This study was followed up by da Costa and van der Schans (2008) who claimed that although the NOMAS has been extensively used since 1993; its reliability had yet to be thoroughly investigated and proven. The purpose of their study was to examine the test-retest and inter-rater reliability of the NOMAS. Seventy-five infants included in the study were born at 26-36 weeks postmenstrual age (PMA), and four observers participated. Inter-rater agreement relating to diagnosis was found to be 'moderate' to 'substantial' (Cohen's κ, between 0.40 and 0.65). Once again, due to the absence of methodological rigor in this study, this research demonstrates suggestive evidence of the inter-rater reliability of the NOMAS.

Nyqvist, Rubertsson, Ewald, and Sjoden developed the preterm infant breastfeeding behaviour scale (PIBBS) in 1996; a visual observation method for assessing preterm breastfeeding competence which was developed in collaboration with observers and mothers. The PIBBS assessment tool would allow for the description of the developmental stages present in preterm infant breastfeeding behaviour by both neonatal personnel and mothers (Nyqvist et al., 1996). Test 1 of inter-rater reliability was undertaken in 24 mother/infant breastfeeding observations. A second test of inter-reliability was applied to 10 preterm infants; employing a revised version of the scale data form and manual. Analysis of inter-observer agreement for nonparametric data was performed by Cohen's kappa for two observers. Analyses of variance were made of parametric data. Inter-rater reliability for test 1 resulted in acceptable agreement between the observers, both in regards to percent of agreement and kappa values; however, a lower level of agreement was observed between each examiner and the mothers. While the two observers achieved kappa values ranging between 0.64 and 1.00, agreement among each observer and mothers ranged between 0.27 and 0.86 and 0.19 and 0.86, respectively (Nyqvist et al., 1996). The inter-observer agreement for test 2 differed depending on the items and method. Percentages of absolute agreement between the two observers ranged from 0.50 and 0.90 and kappa values were seen in values 0.05 up to 0.74. Percentage of agreement among observer A and mothers varied from 0.80 to 1.0,

while kappa values ranged from moderate to almost perfect (0.47 to 1.0). Percentage of agreement among observer B and mothers ranged from 0.60 to 0.90, whereas kappa values varied from fair to substantial (0.29 to 0.74). However, as a result of the lack of methodological rigor in this study, this research presents suggestive evidence of the inter-rater reliability of the PIBBS.

A follow-up study by Nyqvist, Sjoden, and Ewald (1999) strived to further describe the development of preterm infants' behaviour during breastfeeding, until full oral feeding and discharge home was achieved. A prospective, descriptive design was used to study 71 mother-infants pairs, while mothers functioned as data collectors using the PIBBS during observations. A scoring system was added to the scale for purpose of the study. Inter-rater agreement was examined through joint observations of the main author (main observer) and mothers, as well as between the main author and another nurse/research assistant. Seventy simultaneous assessments, one to four per mother-infant pair (n=41), were performed by observer A and mothers. These resulted in excellent agreement for nominal scale items (e.g. rooting, areolar grasp), with kappa values ranging from 0.77-0.94. Twenty-nine simultaneous observations, one to four per mother-infant pair (n=21) were made between observers A and B. These yielded good or excellent agreement for nominal scale items, and kappa values of 0.68-0.84. Narrow confidence intervals and moderate standard deviations were reported for parametric data. Yet, due to the absence of methodological rigor in this study, this research demonstrates suggestive evidence of the inter-rater reliability of the PIBBS.

A systematic review of the psychometric properties of feeding assessment tools used in neonates was carried out by Howe et al. in 2008. The purpose of their study was to comprehensively review and compare the psychometric properties of current clinical feeding assessment tools used in the neonatal population. In total, 941 articles were reviewed. The results indicate that none of the psychometric properties of the seven assessment tool groups identified were satisfactory, and limited representativeness of the samples of the research was observed in all tools. Overall, this study displays suggestive evidence of the inter-rater reliability of both the NOMAS and PIBBS.

Discussion

Appraisal of the Results

Since the body of literature investigating the inter-rater reliability for both the NOMAS and PIBBS is limited, the veracity of the scientific soundness of both instruments is called into question. Many methodological issues need to be taken into consideration when evaluating the evidence.

Participants Selection

Since reliability is based on the "proportion of the total observed variance that is attributable to error", reliability measures will be more accurate as the total variance increases (Portney & Watkins, 2000, p. 559, as cited in Howe et al., 2008). As a result, it is imperative that studies which examine the reliability of an assessment measure include patients that vary in degree of functioning, from normal all the way to severely impaired. None of the NOMAS or PIBBS studies included patients of adequate variability in levels of functioning that can be said to be fully representative of the range from normal to severe.

The NOMAS investigation completed by Palmer and colleagues (1993) excluded infants with structural defects from the sample, and infants who could not complete NOMAS testing due to 'bailout' criteria (e.g. nonnutritive sucking was not observed) were not included for evaluation. Despite the fact that da Costa and van der Schans (2008) attempted to include a more variable sample of participants (i.e., inclusion of a control group of 23 full-term neonates), infants with particular medical conditions were still excluded from the sample (e.g. infants with multiple congential disorders, among others). The excluded patients in both NOMAS studies likely represented a population who may potentially be at most risk for feeding difficulties, and whom would benefit greatly from early identification.

Additional limited sample representations in the studies was also observed in both of the NOMAS studies as they failed to include any breastfeeding infants in their sample, although the assessment measure is intended for both bottle- and breast-feeding observation. In addition, evidence from the literature identifies differences in infants' response to breast- and bottle-feeding (Meier & Anderson, 1987) For example, during breastfeeding infants suck with a more open mouth position, in shorter bursts, with longer pauses, and feeding sessions are of longer length (Meier & Anderson, 1987).

While Palmer et al.'s 1993 NOMAS reliability study at least attempted probability sampling by randomly selecting its participants from infants

meeting inclusion criteria, the two PIBBS studies did not. Instead, investigations by Nyqvist and collaborators (1996; 1999) included a sampling procedure that was based solely on the convenience of the researcher (i.e., a convenience sample). There was no mention in either of the PIBBS articles as to whether or not the researchers made any attempt to make certain that the sample was an accurate representation of the desired population. Instead, the researchers included individuals on the basis of availability. Due to the fact that an unknown portion of the population was excluded, bias is likely present in the convenience sample, and the degree to which the sample is actually representative of the entire population of neonates cannot be discerned (Lunsford & Lunsford, 1995).

Small sample sizes were also noted in three out of the four reviewed studies; n=35 for Palmer et al.'s (1993) NOMAS study, n=24, n=10 and n=41, n=21 for Nyqvist et al.'s 1996 and 1999 studies, respectively. Smaller samples, (i.e., n<30), are less likely to be acceptable representations of population characteristics (Howe et al., 2008), and power is said to be significantly reduced (Portney & Watkins, 2000, as cited in Howe et al, 2008). The NOMAS investigation completed by da Costa & van der Schans (2008) was the only study that included a relatively larger sample size (n=75).

Therefore, small sample sizes in combination with the lack of representativeness of the target population with which they were designed for, limits both the scientific integrity and generalizability of both the NOMAS and PIBBS to the neonatal population.

Procedures

Whether the examiners rating both the NOMAS and PIBBS were blinded to diagnoses of the patients they were rating was also a significant methodological concern. In three out of the four reviewed investigations (Palmer et al., 1993; Nyqvist et al., 1996; Nyqvist et al., 1999), there was no mention as to whether or not the observers possessed any information about the participants, which could have systematically influenced the way in which they administered, scored, or interpreted the results (Dollaghan, 2007). In addition, it should also be noted for the three aforementioned studies, the creators of the experimental tool also served as observers, which further increases the possibility of experimenter bias.

The time that was given to each rater in both NOMAS studies to assess the recorded material was not specified and controlled for (Palmer et al., 1993; da Costa & van der Schans, 2008). Although the NOMAS investigations were carried out in such a way as to simulate a real-life clinical experience, the fact that external raters instead viewed recorded sessions at a later date with no mentioned protocol with regards to time or viewing limitations may have affected examiner evaluation. For example, repeated viewings of the recording may have resulted in a higher rating of a child's feeding skills due to the examiner's increased familiarity of the material.

No standard procedure for time and amount of observations by both examiners and mothers was present in both PIBBS studies (Nyqvist et al., 1996; Nyqvist et al., 1999); assessments took place at any time during the day, according to the mother's convenience (Nyqvist et al., 1996), or as often as mothers could throughout the infant's hospital stay (Nyqvist et al., 1999).

Whereas independent raters were noted in both NOMAS studies (Palmer et al., 1993; da Costa & van der Schans, 2008), the PIBBS investigations (Nyqvist et al., 1996; Nyqvist et al., 1999) utilized joint observation sessions between examiners and mothers on several occasions, which may have affected the level of agreement acquired.

Statistical Analysis

Although Cohen's Kappa statistic (k) and percentages of absolute agreement are commonly used to evaluate item-level agreement (Howe et al., 2007), the Kappa statistic is believed to be the best measure to ascertain agreement between assessors in the case of nominal data analysis, since it takes into account agreement based on chance (Popping, 1983, as cited in da Costa & van der Schans, 2008). Although a k value of more than 0.60 is deemed an acceptable reliability (Tooth & Ottenbacher, 2004), a k value of 0.80 or higher is considered 'almost perfect' or 'satisfactory' (Popping, 1983; Cohen, 1960; Landis & Koch, 1977, as cited in da Costa & van der Schans, 2008). While a percentage of absolute agreement of 80% or more is also considered an acceptable reliability (Brouwer, Reneman, Dijkstra, Groothoff, Schellekens, & Goeken, 2003, as cited in Howe et al., 2008), the measure does not take into account agreement based on chance.

Acceptable reliability was found in Palmer and collaborators' (1996) NOMAS study; however, it should be noted that one of the creators of the tool was also one of the observers, and only percentage

agreement values were reported. In contrast, da Costa and van der Schans' (2008) NOMAS study reported only 'moderate' to 'substantial' inter-rater agreement (i.e., Cohen's k, between 0.40 and 0.65).

The PIBBS studies (Nyqvist et al., 1996; Nyqvist et al., 1999) reported both Kappa statistics and percentages of absolute agreement, as well as correlation coefficients (r) and confidence intervals for parametric data. While consistent adequate inter-rater reliability was found between observers (Nyqvist et al., 1996, Nyqvist et al., 1999), it should be considered that one of the developers of the assessment tool subsequently also served as the main author and observer for both studies.

It was surprising to discover that Nyqvist et al.'s (1999) study was excluded from Howe et al.'s (2008) systematic review, since it undoubtedly fit the authors' inclusion criteria, and would have provided further support for the scientific soundness of the PIBBS. In addition, it should also be known that Howe et al.'s (2008) study did not include a review of da Costa and van der Schans' (2008) NOMAS study since it did not fit inclusion criteria due to its 2008 publishing date. The exclusion of both of the articles in the systematic review may have biased the authors in identifying the NOMAS as being a more psychometrically sound measure in comparison to other assessment tools, since at the time of Howe et al.'s publication, it had more published reports of inter-rater reliability than all others.

Clinical Implications

Although the studies examined (Palmer et al., 1993; da Costa & van der Schans, 2008) indicate the potential usefulness of the NOMAS as a tool for providing detailed observation of an infant's sucking pattern for the purpose of identifying feeding difficulties, it has yet to be proven as an adequate diagnostic tool. This is due to the persistent lack of agreement in scoring separate items, and/or in interpretation of some items belonging to the diagnosis 'disorganization' (Palmer et al., 1993; da Costa & van der Schans, 2008) and to the single reported 'moderate' to 'substantial' inter-observer agreement with respect to diagnosis data (da Costa & van der Schans, 2008).

Since research shows behavioural differences in bottle-versus breast-feeding performance, and neither of the NOMAS studies (Palmer et al., 1993; da Costa & van der Schans, 2008) included breastfeeding infants in their samples, the NOMAS' appropriateness for assessing breast-feeding is still unclear.

Finally, it should be noted that since Palmer et al.'s 1993 study, Palmer has made several revisions to the scale (e.g. added one item to 'dysfunction' category, redefined one item, transferred one item from category 'disorganized' to 'dysfunctional'). Unfortunately, the reliability of this revised version has still not been investigated (da Costa & van der Schans, 2008; Howe et al., 2008).

Examination of inter-rater reliability between observers for the PIBBS has for the most part resulted in acceptable agreement. While initial values of inter-observer agreement between observers and mothers were found to be not quite as satisfactory (Nyqvist et al., 1996), using a revised version of the scale data form and manual and providing an extended period of instruction to mothers appeared to result in a significant increase in inter-rater reliability (Nyqvist et al., 1996; Nyqvist et al., 1999) for both nominal (e.g. rooting, areolar grasp) and measurement data (e.g. duration of latching on). The fact that acceptable agreement can be made between professionals and trained caregivers is very promising for the capability of the PIBBS to be utilized as a breastfeeding assessment tool in the clinical setting. It may potentially be useful during initial breastfeeding sessions, where clinicians can direct mothers to be more responsive to infant behavioural cues, to aid mothers in identifying their infant's emerging feeding skills and competence (Nyqvist et al., 1996, Nyqvist et al., 1999), and incorporate caregiver feedback into intervention planning.

Both the NOMAS and PIBBS are useful clinical assessment tools as they evaluate different aspects of feeding competency. For example, while the NOMAS mainly assesses the biomechanical components for successful feeding, the PIBBS can provide clinicians with information about additional aspects of the feeding process, such as the maternal-infant interaction process or the infant's states during feeding (Howe et al., 2008).

The inception of a 'gold standard', universal rating scale that is used consistently and accurately across the neonatal population appears to be far from a reality. Despite this fact however, current assessment tools such as the NOMAS and the PIBBS have demonstrated their potential usefulness within the clinical setting, as long as clinicians are exceptionally vigilant when interpreting assessment

results, and have an increased awareness of the limitations of visual observational methods, and the number of methodological concerns and lack of proven psychometric soundness for both tools.

Additionally, researchers are strongly encouraged to:
a) Randomly select patients for inclusion in the study that reflect the full range of the neonatal population.
b) Ensure raters are blinded to diagnoses of patients being assessed.
c) Use external raters as observers.
d) Conduct investigations with larger sample sizes.
e) Complete well-designed research studies in order to continue to examine the scientific integrity of the NOMAS and PIBBS instruments related to early identification and evaluation of treatment approaches.

Conclusion

Although there is currently no standard neonatal feeding assessment being used in clinical practice, tools such as the NOMAS or PIBBS are successful in supplying an initial framework for a more organized and systematic means of direct visual observation, which can be valuable to guiding intervention. With more rigorous investigations, these tools will be able to further provide information about the development of the sucking and feeding sequence, and which observable areas are most helpful in predicting and identifying future feeding problems.

References

Aaronson, N., Alonso, J., Burnam, A., Lohr, K. N., Patrick, D. K., Perrin, E., & Stein, R. E. K. (2002). Assessing health status and quality-of-life instruments: attributes and review criteria. *Quality of life research: an international journal of quality of life aspects of treatment, care and rehabilitation, 11*, 193-205.

Arvedson, J. C. (2008). Assessment of pediatric dysphagia and feeding disorders: Clinical and instrumental approaches. *Developmental Disabilities Research Reviews, 14*(2), 118-127.

da Costa, S. P., & van der Schans, C. P. (2008). The reliability of the neonatal oral-motor assessment scale. *Acta Paediatrica (Oslo, Norway: 1992), 97*(1), 21-26.

Dollaghan, C. (2007). *The handbook of evidence-based practice in communication disorders.* Baltimore: Paul H. Brookes Publishing Co.

Hill, P. D., & Johnson, T. S. (2007). Assessment of breastfeeding and infant growth. *Journal of Midwifery & Women's Health, 52*(6), 571-578.

Howe, T. H., Lin, K. C., Fu, C. P., Su, C. T., & Hsieh, C. L. (2008). A review of psychometric properties of feeding assessment tools used in neonates. *Journal of Obstetric, Gynecologic, and Neonatal Nursing: JOGNN / NAACOG, 37*(3), 338-349.

Howe, T. H., Sheu, C. F., Hsieh, Y. W., & Hsieh, C. L. (2007). Psychometric characteristics of the Neonatal Oral-Motor Assessment Scale in healthy preterm infants. *Developmental Medicine & Child Neurology, 49*(12), 915-919.

Leonard, R. & Kendall, K. (1997). *Dysphagia Assessment and Treatment Planning: A Team Approach*. San Diego: Singular Publishing Group, Inc.

Lunsford, T. R., & Lunsford, B. R. (1995). Research Forum: The Research Sample, Part I: Sampling. *Journal of Prosthetics and Orthotics, 7*(3), 105-112.

Meier, P., & Anderson, G. C. (1987). Responses of small preterm infants to bottle-and breast feeding. *The American Journal of Maternal Child Nursing, 12*, 97-195.

Merritt, T. A., Pillers, D., & Prows, S. L. (2003). Early NICU discharge of very low birth weight infants: A critical review and analysis. *Seminars in Neonatology, 8*, 95-115.

Nyqvist, K. H., Rubertsson, C., Ewald, U., & Sjoden, P.O. (1996). Development of the preterm infant breastfeeding behaviour scale (PIBBS): A study of nurse-mother agreement. *Journal of Human Lactation, 12*, 207-219.

Nyqvist, K. H., Sjoden, P.O., & Ewald, U. (1999). The development of preterm infants' breastfeeding behaviour. *Early Human Development, 55*, 247-264.

Palmer, M. M., Crawley, K., & Blanco, I. A. (1993). Neonatal oral-motor assessment scale: A reliability study. *Journal of Perinatology:*

Official Journal of the California Perinatal Association, 13(1), 28-35.

Rogers, B., & Arvedson, J. C. Assessment of infant oral sensorimotor and swallowing function (2005). *Mental Retardation and Developmental Disabilities Research Reviews, 11*(1), 74-82.

Subramaniam, K. (2001). *A qualitative inquiry into the development and use of knowledge in paediatric occupational therapy*. M.Sc. dissertation, University of Toronto, Canada. Retrieved November 20, 2008, from Dissertations & Theses: Full Text database.

Tooth, L. R., & Ottenbacher, K. J. (2004). The kappa statistic in rehabilitation research: An examination. *Archives of Physical Medicine and Rehabilitation, 85,* 1371-1376.