

Original Research

Measurement properties of the Premature Infant Pain Profile-Revised applied at the bedside by physical therapists in the NICU

Amanda dos Santos Erhardt^a, Mariana Bueno^b, Taís Beppler Martins^a,
 Natalia Alves Menegol^a, Dayane Montemezzo^a, Luciana Sayuri Sanada^{a,*}

^a Postgraduate Program in Physical Therapy, Department of Physical Therapy, Universidade do Estado de Santa Catarina, Brazil

^b Lawrence Bloomberg Faculty of Nursing, University of Toronto, Canada



ARTICLE INFO

Keywords:

Neonate
 Pain assessment
 Reliability
 Responsiveness

ABSTRACT

Background: It is essential to accurately assess distress and pain in neonatal intensive care unit (NICU); however, few instruments have had their measurement properties tested for the Brazilian population.

Objective: To analyze the intra- and inter-examiner reliability, internal consistency, and responsiveness of the Brazilian Portuguese version of the Premature Infant Pain Profile-Revised (PIPP-R) scale in Brazilian neonates.

Methods: This is a methodological study conducted in the NICU. Neonates with a gestational age of 24–42 weeks who were not under the effect of muscle block or analgesia at the time of evaluation were included. Inter-examiner reliability was assessed at bedside by two trained evaluators who independently assessed the neonates at bedside and in real time using the PIPP-R. Procedures were filmed and used for intra-examiner reliability assessment after 10–14 days. The Intraclass Correlation Coefficient (ICC) was used to determine intra- and inter-examiner reliability. Responsiveness was assessed by comparing the total scores before and after painful procedures using a paired *t*-test, followed by an effect size analysis.

Results: A total of 119 assessments were performed on 15 neonates. The PIPP-R demonstrated excellent intra- and inter-examiner reliability (ICC > 0.9), and successfully detected changes after an acute painful procedure ($p = 0.003$; effect size = 0.8).

Conclusion: Excellent intra- and inter-examiner reliability, and sensitivity to changes over time were observed by using the PIPP-R at bedside, indicating that this is a suitable instrument for clinical use.

Introduction

In Neonatal Intensive Care Units (NICU), physical therapy has advanced significantly, evolving towards a more comprehensive, neonate-family-centered care approach.^{1–3} Through the accurate identification and appropriate interpretation of signs of discomfort and overstimulation, physical therapists play a critical role in promoting physiological stability and supporting developmental trajectories while minimizing stress to the neonate. Within this framework, the assessment of neonatal pain and stress indicators is a pivotal component to ensure that therapeutic interventions are both effective and consistent with the principles of individualized, developmentally supportive care.^{1–4}

In this context, physical therapists assume a central role in the assessment and modulation of neonatal pain, particularly due to their active involvement in delivering sensory-motor stimulation and respiratory interventions commonly applied in the NICU setting.^{4–6} The use

of validated and culturally adapted pain assessment scales is not only essential for ensuring the accuracy and reliability of clinical evaluations, but also serves to support physical therapists' clinical reasoning and therapeutic decision-making. These tools are instrumental in aligning physical therapy care with the principles of family-centered and developmentally supportive care by enabling individualized pain management strategies and promoting positive neurodevelopmental outcomes.⁷ The systematic review by Nunes et al. highlights that respiratory physical therapy interventions are generally safe and can lead to improvements in pulmonary mechanics and vital parameters without significantly increasing neonatal pain.⁴ This evidence reinforces the importance of integrating sensitive and reliable pain assessment instruments into clinical practice, thereby allowing physical therapists to monitor responses to interventions more precisely and to optimize care for this vulnerable population.

A systematic review identified 65 validated measurement tools for

* Correspondence author: Luciana Sayuri Sanada. Rua Pascoal Simone, 358, Florianópolis-SC, Brazil
 E-mail address: luciana.sanada@udesc.br (L.S. Sanada).

<https://doi.org/10.1016/j.bjpt.2026.101576>

Received 11 November 2024; Received in revised form 9 December 2025; Accepted 13 December 2025

Available online 30 January 2026

1413-3555/© 2026 Associação Brasileira de Pesquisa e Pós-Graduação em Fisioterapia. Published by Elsevier España, S.L.U. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

assessing pain and sedation during the pre-verbal stage of development. However, it was noted that few studies conducted factor analysis or similar methods to select well-defined representative items, and construct validity, internal consistency, and interobserver reliability were not evaluated for all measures.⁸ Menegol et al. identified that out of >40 neonatal pain measurement tools available in the literature, only four have been adequately translated and adapted to Brazilian Portuguese. The authors also noted that the instruments available for the Brazilian population demonstrated low methodological quality as evaluated by the COnsensus-based Standards for the Selection of Health Measurement INstruments (COSMIN), advising caution in interpreting assessments conducted with these scales.⁹ Among the scales translated and adapted to Brazilian Portuguese, the Premature Infant Pain Profile-Revised (PIPP-R) and the Neonatal Infant Pain Scale (NIPS) were developed to assess acute pain.^{8,9} Content validity, reliability, and internal consistency were assessed for the NIPS, whereas only content and construct validity were assessed for the PIPP-R.⁹ The PIPP-R is a revised version of the Premature Infant Pain Profile (PIPP), which was modified to facilitate its clinical use.¹⁰ The PIPP is a reliable and valid measure of acute pain in neonates and several validation studies demonstrate this is a robust measurement tool for neonatal pain.^{8,10-14} For the PIPP-R, the indicators were maintained; however, the scoring method was changed for heart rate, oxygen saturation, facial activity, initial behavioral state, and gestational age.¹⁰

After the translation, content, and construct validation of the PIPP-R into Brazilian Portuguese, we identified the need to further evaluate the measurement properties of this tool. Therefore, this study aimed to assess inter and intra-rater reliability, internal consistency, and responsiveness of the Brazilian version of the PIPP-R.

Methods

Experimental design

This cross-sectional methodological study assessed the measurement properties of the PIPP-R¹⁵ using the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) checklist.¹⁶ The study protocol was approved by the Ethics Committee of Human Research (protocol number: 36,633,320.0.0000.0118). Data were collected in the NICU of a maternity hospital in Florianópolis, Santa Catarina, Brazil.

Participants

Neonates in the NICU were eligible for inclusion if they were clinically stable at the time of evaluation, had chronological or corrected gestational age between 24 and 42 weeks and 6 days, and were scheduled to undergo a routine procedure deemed painful. Neonates receiving muscle blockers or analgesic medications and those on phototherapy were excluded. Parents of eligible neonates admitted to the NICU were approached by a research team member, who explained the study and obtained written consent.

The present study followed the recommended sample size of $n = 50-100$.¹⁷ The sample selection method was non-probabilistic and intentional.

Instruments

The PIPP-R is a multidimensional pain assessment tool comprised of seven items, and it has been widely used to assess acute pain in neonates.¹⁰⁻¹² The tool was translated into Brazilian Portuguese and underwent initial validation. The content validity index attributed by the experts for both clarity and relevance was 1.0. For construct validity, which is based on the correlation between PIPP and PIPP-R scores according to analgesic interventions, the correlation coefficients ranged from $r = 0.96$ to 1.00 ($p < 0.001$).¹³ The PIPP-R includes three behavioral indicators (facial actions: eyebrow bulging, eye squeezing, and

nasolabial furrowing), two physiological indicators (heart rate and oxygen saturation), and two contextual factors (gestational age and behavioral state). Each item is numerically scaled and scored on a 4-point scale (0, 1, 2, and 3), indicating increasing changes in each variable relative to baseline values. The scores obtained for the seven items are summed to deliver the total pain intensity score. The maximum possible score is 21 for premature neonates (<28 weeks of gestational age) and 18 for full-term neonates.^{10,12}

Procedures

All procedures observed were deemed clinically required by the NICU staff. Manipulative techniques for chest physical therapy, nasal suctioning, orotracheal suctioning, insertion/removal of nasogastric or orogastric tubes, changing bandages (central line sticker, transcutaneous oxygen pressure sticker), tracheal intubation, tracheal extubation, venipuncture, arterial puncture, heel puncture, insertion of continuous positive airway pressure, insertion of a central catheter, and insertion or removal of venous access were considered painful procedures in this study.^{18,19}

All neonates were monitored for heart rate and oxygen saturation using the unit's equipment (Vismo PVM-4000 series - Nihon Kohden). For inter-rater reliability, two trained examiners independently assessed pain intensity using the PIPP-R in real-time at bedside (incubator or crib). Examiners recorded all PIPP-R indicators using pen and paper. All examiners were qualified physical therapists with at least 5 years of experience in pediatrics.

A 2-minute video recording of the neonate's face with the camera (iPhone 11 smartphone) positioned at the top of the incubator or crib, at an approximate distance of 36 to 43 cm from the neonate's face was taken. One of the examiners re-assessed the behavioral state and facial actions after 10-14 days of the procedure. The examiner watched the video once, in real time. The same physiological indicators obtained at bedside were used. Scores observed in real time and obtained from the recordings were compared to assess intra-examiner reliability.

The responsiveness of the PIPP-R was assessed by comparing the pain intensity scores before (baseline) and after painful procedures, by the same physical therapist researcher. These painful procedures were performed by the healthcare staff, as described above and they decided on the need to use non-pharmacological pain management techniques for the procedures, according to local practices; this information was not recorded as part of the study.

Data analysis

Statistical analyses were performed using IBM SPSS 20.0 (IBM Corp., Armonk, NY, USA). The variables are presented using measures of central tendency, dispersion, 95 % confidence intervals (95 % CI), and absolute and/or relative frequency. The normal distribution of the data was assessed using the Kolmogorov-Smirnov test. The significance level was set at 0.05.²⁰ Internal consistency was evaluated using Cronbach's alpha, with acceptable values defined as ≥ 0.7 .²¹

To analyze inter-rater reliability, the Mann-Whitney test compared the assessments of the two examiners, and the Intraclass Correlation Coefficient (ICC) was applied using two-way random effects, absolute agreement, and mean measurement models.²²⁻²⁴ For intra-rater reliability, the Wilcoxon test was used to compare the two assessments (real-time and video), and the ICC was applied using two-way mixed effects, absolute agreement, and mean measurement models.²² Reliability values were categorized as follows: ICC > 0.9 (excellent); ICC 0.75-0.9 (good); ICC 0.5-0.75 (moderate); and ICC < 0.5 (poor reliability).²² The standard error of measurement (SEM) was calculated by using the formula $SEM = \sqrt{(\sigma_0^2 + \sigma_{residual}^2)}$, where σ_0^2 represents the variance owing to systematic differences between the examiner and $\sigma_{residual}^2$ represents the random error variance.^{23,24}

The correlation between the subtotal and total score variables

between the intra- and inter-rater assessments was analyzed using the Spearman correlation test. The correlation values were interpreted as follows: ≤ 0.25 (weak or no correlation); $0.25-0.5$ (regular correlation); $0.5-0.75$ (moderate to good correlation); and > 0.75 (good to excellent correlation).²⁰ Bland-Altman analysis was employed to verify the agreement between the intra- and inter-rater reliability assessments. This analysis was based on the mean and standard deviation of the differences, as well as the lower and upper limits of agreement. In addition, the error (defined as the dispersion of the difference points around the mean) was analyzed. To confirm agreement between the scores in the Bland-Altman analysis, the bias (range of differences between the averages deviating from zero), and the error (dispersion of the difference in points around the average) must be examined, given that at least 90–95 % of the measurements must fall within the limit of agreement.^{25–27} Lin's Concordance Correlation Coefficient (CCC) was used for the analysis of agreement by ascertaining the magnitude of deviation from the line of perfect agreement.^{28,29} To assess the degree of agreement by CCC, the following classification was used: very good agreement (0.81 to 1), good agreement (0.61 to 0.80), moderate agreement (0.41 to 0.60), fair agreement (0.21 to 0.40), and poor agreement (< 0.2).^{30,31}

Responsiveness was assessed by comparing total scores before and after the painful procedure. The analysis was performed using the paired *t*-test, along with mean difference (95 % CI) and the effect size analysis, which considered the difference in means divided by the standard deviation of the mean at time zero (the duration from the first assessment to the time before the painful procedure) between moments. The effect size was interpreted as small at 0.20, medium at 0.50, and large at 0.80.²⁰

Results

A total of 119 assessments were performed on 15 neonates (10 males), and the median gestational age at birth was 28 weeks, with an interquartile range of 4 weeks. The PIPP-R demonstrated an overall internal consistency of 0.753 using Cronbach's alpha.

The intra-rater reliability and SEM data are presented in Table 1. A comparative analysis between the two intra-rater reliability assessments for the PIPP-R did not reveal statistically significant differences across all items ($p > 0.05$), indicating no detectable differences within the limits of this analysis. Inter-rater reliability and SEM data are presented in Table 2. Similarly, no statistically significant differences were found between the two raters for all items ($p > 0.05$), indicating that potential differences, if any, were not captured by the current study design. Excellent reliability was observed for both intra- and inter-rater assessments (ICC = 0.98 and 0.90, respectively). Fig. 1 shows the data analysis using Lin's Concordance Correlation Coefficient.

The qualitative analysis of agreement between the PIPP-R subtotal and total score data was conducted using the Bland-Altman analysis (Fig. 2). The intra-rater bias was close to zero, both for the subtotal and total score analyses of PIPP-R, with no statistically significant

differences [(95 % CI: $-0.11, 0.13, p = 0.89$) and (95 % CI: $-0.12, 0.17, p = 0.74$), respectively], indicating agreement. In the linear regression analysis for the intra-rater comparison, no significant proportional bias for the subtotal and total scores was observed, ($p = 0.98$ and 0.91 , respectively). When analyzing the inter-rater bias of the subtotal and the total PIPP-R, values close to zero and no statistically significant differences were observed, [(95 % CI: $-1.25, 0.42, p = 0.28$ and (95 % CI: $-0.25, 0.35, p = 0.73$), respectively]. Furthermore, in the linear regression analysis for the inter-examiner comparison of the subtotal and total scores, no significant proportional bias was observed ($p = 0.25$ and 0.25 , respectively).

Regarding the responsiveness of the PIPP-R, there was a significant capacity for change on the score after an acute painful procedure compared to baseline with a mean difference of 1.5 points (95 % CI: $0.42, 1.15; p < 0.01$; effect size = 0.8).

Discussion

This study assessed inter- and intra-examiner reliability, internal consistency, and responsiveness of the Brazilian version of the PIPP-R. The results indicated excellent reliability, demonstrated by high intra- and inter-examiner agreement; and strong internal consistency, demonstrated by excellent correlation of the indicators. Furthermore, the scale was sensitive to changes in acute pain in neonates. The identification of reliable assessment tools helps strategize neonatal pain management and improve the quality of care for vulnerable newborns.³² Preterm and sick newborns are commonly exposed to a large number of painful procedures during their stay in the NICU, which increases the risk of negative neurobiological effects.³³ Evidence-based practice combined with objective assessments of the value and effectiveness of rehabilitation assessment and treatment techniques should be prioritized.³⁴ However, available tools in Brazilian Portuguese to measure neonatal pain are scarce. Of the > 40 neonatal pain assessment instruments available in the literature, only four have been formally translated and adapted for this population.⁹

Construct and content validity of the Brazilian version of the PIPP-R were previously analyzed in a study that demonstrated high correlation between pain scores for PIPP and PIPP-R for procedures using different pain relief strategies as well as for different types of procedures in both full-term and preterm neonates.¹³ Strong internal consistency was also demonstrated. Measurement tools with strong internal consistency typically exhibit correlations ranging from 0.70 to 0.90 between their items. Lower correlations (< 0.70) indicate that the tool's items are possibly measuring different phenomena. Conversely, a high correlation (> 0.90) indicates redundant items, which may limit the content validity of the scale.²⁰ Consistent with our findings, a systematic review of the validity and reliability of the instruments used to assess behavior, stress, and/or pain in preterm newborns in the NICU revealed that the internal consistency of PIPP-R, measured by the Cronbach's alpha, varied between 0.71 to 0.84.¹⁴

In addition, this study demonstrated that the PIPP-R has excellent

Table 1

Intra-examiner reliability of *Premature Infant Pain Profile-Revised* (PIPP-R) ($n = 119$).

	Mean (SD) Bedside	Mean (SD) Video	p^a	Mean (SD)	ICC	95 % CI		p^b	SEM
						LB	UB		
Brow bulge	1.12 (1.3)	1.09 (1.3)	0.36	1.10 (1.3)	0.99	0.981	0.991	< 0.01	0.29
Eye squeeze	0.54 (0.9)	0.57 (1.0)	0.43	0.55 (0.3)	0.92	0.886	0.945	< 0.01	0.46
Nasolabial furrow	0.18 (0.6)	0.16 (0.6)	0.16	0.17 (0.1)	0.99	0.985	0.993	< 0.01	0.16
Baseline BS	1.62 (1.2)	1.61 (1.2)	0.79	1.61 (0.1)	0.96	0.948	0.975	< 0.01	0.34
Subtotal	2.77 (2.6)	2.76 (2.6)	0.98	2.77 (0.1)	0.98	0.975	0.998	< 0.01	0.49
Total Score	5.66 (2.8)	5.64 (2.8)	0.86	5.65 (0.2)	0.98	0.971	0.986	< 0.01	0.60

BS, behavioural state; CI, confidence interval; HR, heart rate; ICC, intraclass correlation coefficient; LB, Lower Bound; SatO₂, oxygen saturation; SD, standard deviation; SEM, standard error of measurement; UB, Upper Bound

^a p value of Wilcoxon Test

^b p value of ICC.

Table 2
Inter-examiner reliability of *Premature Infant Pain Profile-Revised* (PIPP-R) (n = 95).

	Mean (SD) Examiner 1	Mean (SD) Examiner 2	p ^a	Mean (SD)	ICC	95 % CI		p ^b	SEM
						LB	UB		
Change in HR	0.61 (0.7)	0.60 (0.7)	1.00	0.59 (0.1)	0.96	0.947	0.977	<0.01	0.20
Decrease in SatO ₂	0.38 (0.6)	0.38 (0.6)	0.82	0.38 (0.1)	0.93	0.899	0.955	<0.01	0.23
Brow bulge	0.98 (1.2)	0.90 (1.1)	0.80	0.93 (0.7)	0.88	0.822	0.921	<0.01	0.75
Eye squeeze	0.44 (0.8)	0.43 (0.8)	0.91	0.42 (0.1)	0.80	0.700	0.867	<0.01	0.47
Nasolabial furrow	0.13 (0.5)	0.07 (0.4)	0.36	0.10 (0.5)	0.90	0.845	0.932	<0.01	0.41
Baseline BS	1.74 (1.2)	1.84 (1.2)	0.57	1.80 (0.9)	0.91	0.869	0.942	<0.01	0.82
Subtotal	2.53 (2.3)	2.38 (2.1)	0.78	2.42 (0.1)	0.90	0.853	0.935	<0.01	1.39
Total Score	5.10 (2.6)	5.04 (2.4)	0.90	5.06 (0.1)	0.90	0.855	0.936	<0.01	1.11

BS, behavioural state; CI, confidence interval; HR, heart rate; ICC, intraclass correlation coefficient; LB, Lower Bound; SatO₂, oxygen saturation; SD, standard deviation; SEM, standard error of measurement; UB, Upper Bound.

^a p value of Mann-Whitney Test.

^b p value of ICC.

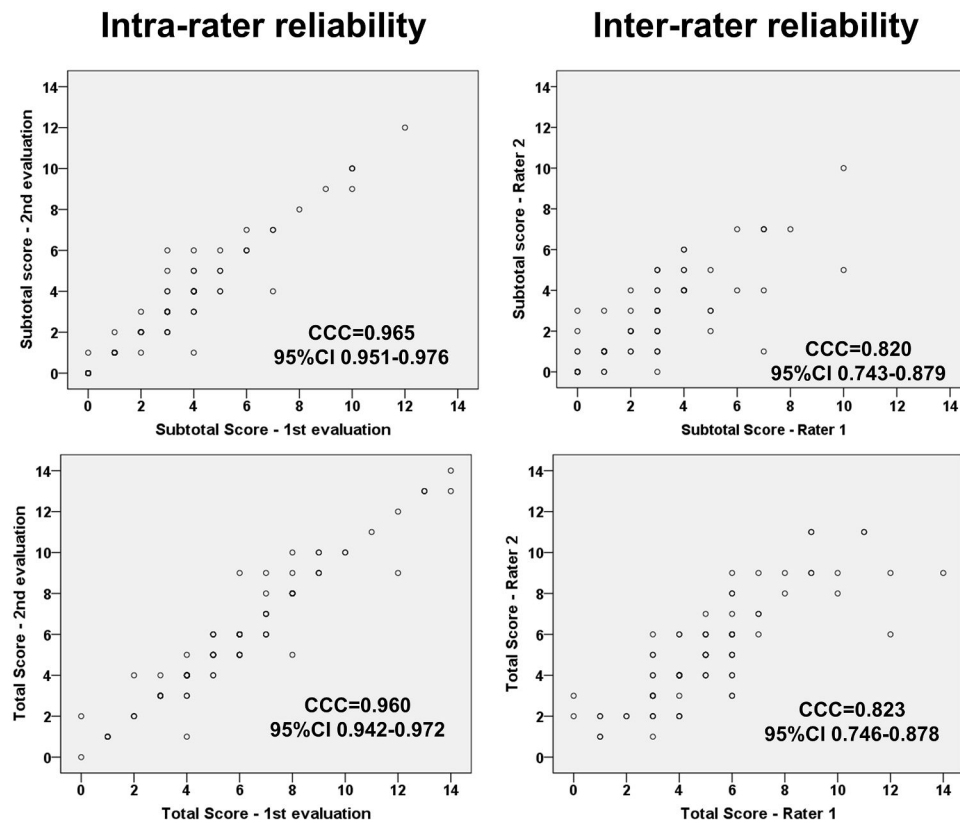


Fig. 1. – Lin's concordance correlation coefficient: intra- and inter-examiner of subtotal and total score of PIPP-R.

intra-examiner reliability. To our knowledge, studies on the intra-examiner reliability of the PIPP-R are scarce. Among the studies evaluating the measurement properties of the PIPP, intra-examiner reliability was analyzed in the study by Ballantyne et al., who reported excellent reliability for the PIPP scale (ICC = 0.94–0.98).³⁵

In the present study, inter-examiner reliability was excellent, corroborating the findings of previous studies.^{12,36–38} Taplak and Bayat assessed inter-examiner reliability using video recordings and reported excellent reliability (ICC = 0.94–1.00).³⁶ In line with this study's methodology, in which two independent examiners rated pain scores at bedside and in real-time, studies reporting psychometric properties of the Persian and Indonesian versions of the PIPP-R also demonstrated excellent inter-examiner reliability (ICC = 0.98–0.99 and 0.97, respectively).^{37,38} By demonstrating intra-examiner reliability of the tool in the 'real world' (e.g., pain assessment at bedside, in real-time) this study helps to advance pain measurement practices in the NICU and to

improve confidence among NICU professionals in using this particular tool.

An agreement analysis is recommended to complement the reliability assessment.²⁷ Our findings demonstrated a good agreement in CCC between the intra- and inter-examiner subtotal and total scores. In this study, the Bland–Altman analysis demonstrated that the mean differences for both intra- and inter-examiner scores, in both subtotal and total scores, remained close to zero without proportion of bias, indicating strong intra- and inter-examiner agreement. While the graphs showed that most values were within the limits of agreement, the clinical interpretation of these limits should guide the assessment of agreement.²⁶ Elias et al., in their study, aimed to analyze whether parents and healthcare professionals homogeneously assessed the presence and magnitude of pain in critically ill neonates. They reported agreement for the Bland–Altman analysis when no pain was present, while disagreement between observers was noted for moderate pain.³⁹ In the

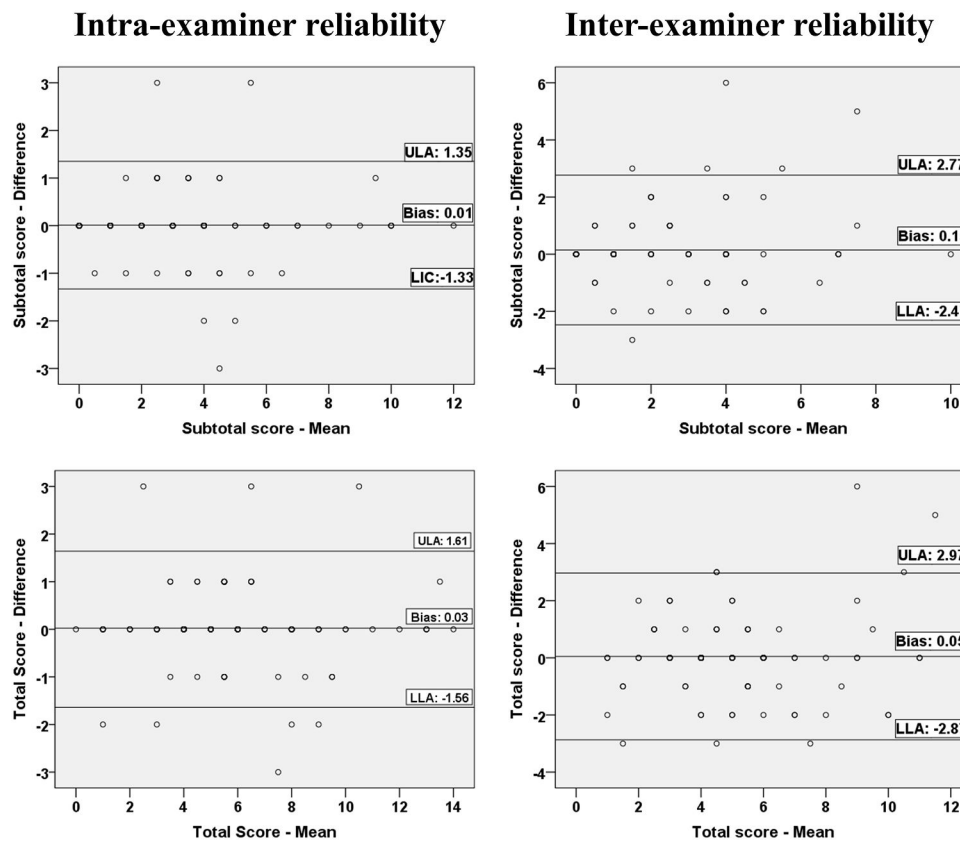


Fig. 2. –Bland-Altman analysis: intra- and inter-reliability of subtotal and total score of PIPP-R.

present study, points outside the limit of agreement for inter-examiner reliability occurred in moderate to severe pain scores.

Our study findings elucidated that PIPP-R exhibited substantial responsiveness to changes after an acute painful procedure. The initial validation of the original version of the PIPP-R demonstrated the instrument's responsiveness between types of events and interventions, revealing a significant difference between PIPP-R scores in non-nutritive sucking with sucrose [PIPP-R 6.4 (3.4)], restraint facilitated with non-nutritive suction and sucrose [PIPP-R 7.2 (3.3)], and non-nutritive sucking alone [PIPP-R 8.6 (4.0)], suggesting greater efficacy in non-nutritive sucking with sucrose intervention.¹¹ However, despite the considerable increase in the number of studies on neonatal pain and neonatal pain assessment, our knowledge of the responsiveness of pain tools is limited.⁴⁰

In our study, pain was assessed using the PIPP-R in real-time, both at bedside (real-time) and through analysis of video recordings (without pauses). The results confirm that the tool is reliable and feasible for implementation in the clinical setting. The PIPP-R may assist physical therapists in recognizing signs of stress or pain in neonates, which may reflect a loss of self-regulation and suggest that continued interaction with physical therapy may not be appropriate at that moment. Such disruptions and observable disorganization highlight the infant's developmental challenges and indicate areas of vulnerability under certain environmental or care conditions.^{3,41}

Additional research on neonatal pain assessment is warranted to facilitate adequate pain management in newborns, and to disseminate the results already obtained. This study provides substantial contributions to this area by analyzing the measurement properties of the PIPP-R, thereby ensuring its safety and accuracy in assessing acute pain in the NICU and serving as a reliable and responsive tool for clinical and scientific use in Brazilian neonates. Despite these potential advantages, our study had certain limitations, including (1) the observation of only 15 neonates; (2) the exclusive use of physical therapists to apply the PIPP-R;

and its cross-sectional design. Future studies should consider employing the Brazilian version of the PIPP-R in research exploring neonatal pain assessment and management.

Conclusion

The Brazilian version of the PIPP-R demonstrated excellent intra- and inter-examiner reliability, strong internal consistency, and good responsiveness in assessing procedural pain in neonates. Therefore, it is feasible to use this tool for procedural pain assessment in the NICU.

Declaration of competing interest

The authors declare no competing interest.

Acknowledgements

Grant sponsor PROAP-AUXEP (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior); FAPESC (Fundação de Amparo à Pesquisa do Estado de Santa Catarina); PAEX-PROCEU/UEDESC n° 01/2021

References

- Byrne E, Garber J. Physical therapy intervention in the neonatal intensive care unit. *Phys Occup Ther Pediatr*. 2013;33(1):75–110. <https://doi.org/10.3109/01942638.2012.750870>.
- Técnico-Científico P. *Parecer Técnico-Científico*. <https://abrafin.org.br/https://assobrafir.com.br/>.
- Sweeney JK, Heriza CB, Blanchard Y, Dusing SC. Neonatal physical therapy. Part II: practice frameworks and evidence-based practice guidelines. *Pediatr Phys Ther*. 2010;22(1):2–16. <https://doi.org/10.1097/PEP.0b013e3181cda43>.
- de Moraes Nunes A, Nascimento Sales Figueiredo Fernandes AT, Tévina de Castro Silva A, Guerra Azevedo I, Souza Monteiro K, Pereira SA. Effects of respiratory physiotherapeutic interventions on pulmonary mechanics, vital parameters and pain in newborns: a systematic review. *Can J Respir Ther*. 2025;61. <https://doi.org/10.29390/001c.140878>.

5. Ribeiro AL, Costa MFP, Silva PYF, et al. Effects of the use of a cocoon on the autonomic, motor, and regulatory systems in preterm newborns: randomized clinical trial. *Arch Pediatr*. 2024;31(4):250–255. <https://doi.org/10.1016/j.arcped.2024.01.005>.
6. Johnston C, Stopiglia MS, Ribeiro SNS, Baez CSN, Pereira SA. First Brazilian recommendation on physiotherapy with sensory motor stimulation in newborns and infants in the intensive care unit. *Rev Bras Ter Intensiva Assoc Med Intensiva Bras - AMIB*. 2021;33(1):12–30. <https://doi.org/10.5935/0103-507X.20210002>.
7. Hodgson CR, Mehra R, Franck LS. Infant and Family outcomes and experiences related to Family-centered care interventions in the NICU: a systematic review. *Child Multidiscip Digit Publ Inst (MDPI)*. 2025;12(3). <https://doi.org/10.3390/children12030290>.
8. Giordano V, Edobor J, Deindl P, et al. Pain and sedation scales for neonatal and pediatric patients in a preverbal stage of development: a systematic review. *JAMA Pediatr Am Med Assoc*. 2019;173(12):1186–1197. <https://doi.org/10.1001/jamapediatrics.2019.3351>.
9. Menegol NA, Ribeiro SNS, Okubo R, Sonza A, Montemuzzo D, Sanada LS. Quality assessment of neonatal pain scales translated and validated to Brazilian Portuguese: a systematic review of psychometric properties. *Pain Manag Nursin*. 2022;23(4): 559–565. <https://doi.org/10.1016/j.pmn.2021.12.003>.
10. Stevens BJ, Gibbins S, Yamada J, et al. The premature infant pain profile-revised (PIPP-R). *Clin J Pain*. 2014;30(3):238–243. <https://doi.org/10.1097/ajp.0b013e3182906aed>.
11. Stevens B., Johnston C., Taddio A., Gibbins S., Yamada J. *The premature infant pain profile: evaluation 13 years after development.*; 2010. doi:10.1097/AJP.0b013e3181ed1070.
12. Gibbins S, Stevens BJ, Yamada J, et al. Validation of the premature infant pain profile-revised (PIPP-R). *Early Hum Dev*. 2014;90(4):189–193. <https://doi.org/10.1016/j.earlhumdev.2014.01.005>.
13. Bueno M, Moreno-Ramos MC, Forni E, Kimura AF. Adaptation and initial validation of the premature infant pain profile-Revised (PIPP-R) in Brazil. *Pain Manag Nurs*. 2019;20(5):512–515. <https://doi.org/10.1016/j.pmn.2019.02.002>.
14. Glenzel L, do Nascimento, Oliveira P, Marchi BS, Ceccon RF, Moran CA. Validity and reliability of pain and behavioral scales for preterm infants: a systematic review. *Pain Manag NursWB Saunders*. 2023;24(5):e84–e96. <https://doi.org/10.1016/j.pmn.2023.06.010>.
15. Hulley S., Cummings S., Browner W., Grady D., Newman T. *Delineando a Pesquisa Clínica*. 4th ed. Artmed; 2015.
16. Kottner J, Audigé L, Brorson S, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *J Clin Epidemiol*. 2011;64(1):96–106. <https://doi.org/10.1016/j.jclinepi.2010.03.002>.
17. Mokkink LB, de Vet H, Diemeer S, Eekhout I. Sample size recommendations for studies on reliability and measurement error: an online application based on simulation studies. *Health v Outcomes Res Methodol*. 2023;23(3):241–265. <https://doi.org/10.1007/s10742-022-00293-9>.
18. Simons SHP, Van Dijk M, Kanwaljeet, et al. Do we still hurt newborn babies? *Prospect Study Proced Pain Analg Neonates*. 2003.
19. Roofthoof DWE, Simons SHP, Anand KJS, Tibboel D, Van Dijk M. Eight years later, are we still hurting newborn infants? *Neonatology*. 2014;105(3):218–226. <https://doi.org/10.1159/000357207>.
20. Portney LG. *Foundations of Clinical Research: Applications to Evidence-Based Practice*. 4th ed. Philadelphia, PA: F.A. Davis Company; 2020.
21. de Souza AC, NMC Alexandre, Guirardello E de B. Propriedades psicométricas na avaliação de instrumentos: avaliação da confiabilidade e da validade. *Epidemiol v Saude*. 2017;26(3):649–659. <https://doi.org/10.5123/S1679-49742017000300022>.
22. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2):155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>.
23. Mokkink LB, Eekhout I, Boers M, van der Vleuten CP, de Vet HC. Studies on reliability and measurement error of measurements in medicine – From design to statistics explained for medical researchers. *Patient Relat Outcome Meas*. 2023;14: 193–212. <https://doi.org/10.2147/prom.s398886>.
24. Veet H.C.W., Terwee C.B., Mokkink L.B., Knol D. *Measurement in medicine: a practical guide*; 2011.
25. Woodfield HC, Gerstman BB, Olaisen RH, Johnson DF. Interexaminer reliability of supine leg checks for discriminating leg-length inequality. *J Manip Physiol Ther*. 2011;34(4):239–246. <https://doi.org/10.1016/j.jmpt.2011.04.009>.
26. Correlation Bunce C. Agreement, and Bland-Altman analysis: statistical analysis of method comparison studies. *Am J Ophthalmol*. 2009;148(1):4–6. <https://doi.org/10.1016/j.ajo.2008.09.032>.
27. Hirakata VN, Camesy SA. Análise de concordância entre métodos de Bland-Altman. *Rev HCPA Fac Med Univ Fed Rio Gd Sul*. 2009;29(3):261–268.
28. Lin L, Hedayat AS, Sinha B, Yang M. Statistical methods in assessing agreement: models, issues, and tools. *J Am Stat Assoc*. 2002;97(457):257–270. <https://doi.org/10.1198/016214502753479392>.
29. Lin LIK. A concordance correlation coefficient to evaluate reproducibility. <https://www.jstor.org/stable/2532051>; 1989.
30. Wadoux AMJC, Minasny B. Some limitations of the concordance correlation coefficient to characterise model accuracy. *Ecol Inf*. 2024;83. <https://doi.org/10.1016/j.ecoinf.2024.102820>.
31. Atkinson G, Nevill A. Comment on the use of concordance correlation to assess the agreement between two variables. <https://www.jstor.org/stable/2533978>; 1997.
32. Oliveira NRG, Formiga CKMR, Ramos BA, Noleto R dos S, Moreira NN, de S, Amaral WN. Confiabilidade e consistência interna na avaliação da dor neonatal de prematuros durante o procedimento de aspiração traqueal. *Braz J Pain*. 2022;5(4): 341–346. <https://doi.org/10.5935/2595-0118.20220065-en>.
33. McPherson C, Miller SP, El-Dib M, Massaro AN, Inder TE. The influence of pain, agitation, and their management on the immature brain. *Pediatr ResSpringer Nat*. 2020;88(2):168–175. <https://doi.org/10.1038/s41390-019-0744-6>.
34. Gadotti I, Vieira E, Magee D. Importance and clarification of measurement properties in rehabilitation. *Rev Bras Fisioter*. 2006;10(2):137–146. <https://doi.org/10.1590/s1413-35552006000200002>.
35. Ballantyne M, Stevens B, McAllister M, Dionne K, Jack A. Validation of the premature infant pain profile in the clinical setting. *Clin J Pain*. 1999;15(4): 297–303.
36. Taplak AŞ, Bayat M. Psychometric testing of the Turkish version of the premature infant pain profile revised-PIPP-R. *J Pediatr Nurs*. 2019;48:e49–e55. <https://doi.org/10.1016/j.pedn.2019.06.007>.
37. Fitri SYR, Lusmilasari L, Juffrie M. The Indonesian version of the Premature Infant Pain Profile-Revised: translation and adaptation of a neonatal pain assessment. *Int J Nurs Sci*. 2019;6(4):439–444. <https://doi.org/10.1016/j.ijnss.2019.06.010>.
38. Sadeghi A, Rassouli M, Abolhasan Gharehdaghi F, et al. Validation of the Persian version of premature infant pain profile-revised in hospitalized infants at the neonatal intensive care units. *Iran J Pediatr*. 2017;27(5). <https://doi.org/10.5812/ijp.10056>.
39. Elias LSDT, Guinsburg R, Peres CA, Balda RCX, Dos Santos AMN. Discordância entre pais e profissionais de saúde quanto à intensidade da dor no recém-nascido criticamente doente. *J Pediatr (Rio J)*. 2008;84(1):35–40. <https://doi.org/10.2223/JPED.1748>.
40. Meesters N, Dilles T, Simons S, van Dijk M. Do pain measurement instruments detect the effect of pain-reducing interventions in neonates? A systematic review on responsiveness. *J PainChurchill Livingstone Inc*. 2019;20(7):760–770. <https://doi.org/10.1016/j.jpain.2018.12.005>.
41. Blanchard Y, Oberg GK. Physical therapy with newborns and infants: applying concepts of phenomenology and synactive theory to guide interventions. *Physiother Theory Pr*. 2015;31(6):377–381. <https://doi.org/10.3109/09593985.2015.1010243>.