The Neurologic and Adaptive Capacity Score Is Not a Reliable Method of Newborn Evaluation

Stephen H. Halpern, M.D., M.Sc., F.R.C.P.C.,* Judith A. Littleford, M.D., B.Sc., F.R.C.P.C.,†
Nicole J. Brockhurst, R.N., M.Sc.,‡ Paul J. Youngs, B.M., F.R.C.A.,§ Nariman Malik, B.Sc.,‖
Holly C. Owen, R.N., B.Sc.N.#

Background: The Neurologic and Adaptive Capacity Score (NACS) is a multi-item scale that was published in 1982 to measure the effects of intrapartum drugs on the neonate. Although this scoring system has been widely used in obstetric anesthesia research, studies confirming its reliability have not been published. The purpose of this study was to assess the reliability of the NACS.

Methods: Two teams of observers were trained to perform the NACS on healthy, term neonates born in the vertex presentation. Two examinations were performed on each neonate within the first 2.5 h of life. Simultaneous (or "split-half") reliability was assessed using the α coefficient. Test–retest reliability was assessed using the intraclass correlation coefficient. The test was considered to be reliable if α was greater than 0.7 and the intraclass correlation coefficient was greater than 0.6.

Results: Two hundred babies were studied. The α was 0.47 and the intraclass correlation coefficient was 0.38 (95% confidence interval, 0.24–0.52).

Conclusions: The NACS had poor reliability both on simultaneous testing and in the test–retest situation when used to evaluate term, healthy neonates. The authors suggest that other measures need to be developed to evaluate the effect of intrapartum drug administration in the neonate. Health measurement scales should undergo rigorous assessment for reliability and validity before they are used in clinical practice or for research purposes.

THE Neurologic and Adaptive Capacity Score (NACS) was developed to detect central nervous system depression in term neonates exposed to intrapartum medications.[1] It consists of 20 maneuvers arranged into two subscales: adaptive capacity and neurologic evaluation. The neurologic subscale is further divided into four parts testing passive tone, active tone, primary reflexes, and general neurologic status. Each of the 20 maneuvers is assigned 0, 1, or 2 points depending on the infant's response. The maximum total score possible is 40.

The impetus for developing the NACS was, in part, to satisfy Food and Drug Administration requirements that drugs used in the clinical arena have a minimal effect on neonatal neurobehavior.[1] The Food and Drug Administration wanted a means of quantifying the neonatal effects of new drugs and established medications. Whether the NACS is a suitable method to satisfy these requirements has not been confirmed.

At the time the NACS was published in 1982, two other scales were widely used by obstetric anesthesiologists to assess neurobehavior in neonates (the Early Neonatal Neurobehavioral Scale[2] and the Brazelton Neonatal Behavioral Assessment Scale[3]). These scales were time-consuming and contained some aversive stimuli. Maneuvers included in the neurologic portion of the NACS promised to be easy to complete (60–90 s) and reproducible. No special equipment was required. The adaptive capacity components entailed closer evaluation, and the responses were subject to interpretation. This portion of the test could take several minutes to perform.

The initial publication described the original observations for 61 term babies who were evaluated at 15 min and 2 and 24 h of age. All babies were born vaginally to mothers who had received a variety of intrapartum medications. The authors reported an interrater reliability of 92.8% when two trained individuals simultaneously observed each NACS examination.

An editorial accompanying publication of the NACS condemned the development of the scale and questioned its validity, both on methodologic grounds and because of inadequacies in the conceptual framework.[4] A second editorial conceded that the reviewers could not agree on the merit of the NACS. Because of this controversy, the editorial went on to say "additional studies will be required to determine the validity and sensitivity of their neonatal examination."[5]

The NACS has been used as an important tool to evaluate neonatal outcome in more than 70 studies comparing intrapartum drugs and other interventions.[6] Apart from the original description of the scale, only one publication has considered the issue of reliability. The study included only five babies and used interobserver reliability to check the performance of NACS testers before starting an unrelated investigation.[7] Another study,

* Director of Obstetrical Anesthesia, Department of Anesthesia, Sunnybrook and Women's College Health Sciences Centre, Associate Professor of Anesthesia, Department of Obstetrics and Gynecology, University of Toronto. † Director of Obstetrical Anesthesia, Department of Anesthesia, Mount Sinai Hospital and the University Health Network, Assistant Professor of Anesthesia, Department of Anesthesia, University of Toronto. ‡Research Nurse, § Fellow, Department of Anesthesia, Mount Sinai Hospital and the University Health Network. ‖ Medical Student, University of Toronto. #Research Nurse, Department of Anesthesia, Sunnybrook and Women's College Health Sciences Centre.

958

which is available in abstract form,[8] failed to show adequate test–retest reliability between observers.

Similarly, there are scant data on the validity of the NACS. The original research showed a high correlation between the NACS and previously developed neurobehavioral scales. This is not surprising because the scales share many items. There have been no studies specifically designed to validate the NACS by correlating it with, for example, cord blood opioid, sedative levels, or breast-feeding outcomes. The purpose of this study was to determine whether the NACS is a reliable tool for measuring neurobehavior in the term neonate.

Methods

The study was approved by the Research Ethics Board at the University of Toronto, and informed consent was obtained from the parents of each neonate enrolled.

Two teams of observers underwent a 2-month training period before beginning the study. During this time, we assessed the description of each item from the original article. Where ambiguity existed, we sought additional information from an earlier publication.[9] Once we had agreed to the descriptions of all the items by consensus, we prepared a training manual that contained photographs and detailed instructions about the technique of examination and the score to be assigned to a given response. Observers trained together by repeated observation and performance of NACS examinations. Differences in technique or scoring were resolved using video demonstration and a baby mannequin. Training continued until all observers could perform and score the test similarly. Midway through the study, all observers rechecked their ability to perform and score the NACS.

The study was conducted at two sites with two separate teams of observers. Seven people were trained: two staff anesthesiologists, two obstetric anesthesia fellows, two research nurses, and one medical student. One observer from each site (a research nurse) was chosen to rate all infants at that site. The second observer was chosen from a pool of the remaining investigators at that site. Neonates were eligible for the study if they were healthy, 38–42 weeks' gestation, of vertex presentation, and born to healthy mothers. We included medicated and unmedicated births, vaginal and cesarean deliveries, and examined all neonates shortly after birth to maximize the spread of possible NACS results.

The first NACS was performed within 2 h of birth. After 30 min, a second tester, blind to the first result, repeated the examination. We chose the interval to be short enough to reduce the likelihood of the infant's neurobehavior changing but long enough for them to recover from the initial examination. The state of the baby (asleep, drowsy, quiet–alert, crying) and time of last feeding were recorded before each test. We examined the infant in similar temperature, sound, and lighting conditions both times. Babies were excluded from analysis if a significant intervention occurred between observations (*e.g.*, changed room, invasive diagnostic procedure).

Statistical Analysis

We used Minitab version 12.22 (Minitab Inc., State College, PA) for all statistical calculations. The data were pooled between sites. The primary outcome was reliability. Two aspects of reliability were studied: simultaneous reliability (also known as internal consistency or split-half reliability) and test–retest reliability. These were assessed using the α coefficient (see below)[10] and the intraclass correlation coefficient (ICC) for the NACS between the two observers, respectively. Secondary outcomes included the ICC for the adaptive capacity subscale and the neurologic examination subscale, as well as the ICC for the NACS of babies who were in the same state for both examinations. In addition, the ICC for the NACS at each site was determined.

Descriptive statistics (mean, median, SD, and range) were performed on all total NACS scores and separately for the two subscales and for babies in the same state for both examinations.

Demographics

Demographic data included maternal age and parity, gestational age, and birth weight. Data concerning the delivery included mode of delivery, Apgar score at 1 and 5 min, and need for resuscitation at birth. We documented the drugs that the fetus was exposed to *in utero* and recorded the cumulative dose of each.

Simultaneous Reliability

The α coefficient was calculated to assess simultaneous reliability of the total scale and the adaptive and neurologic components using the following formula :

$$\alpha = \left(\frac{n}{n-1}\right)\left(1 - \frac{\Sigma \sigma_i^2}{\sigma_T^2}\right) \quad (1)$$

where α = the measure of simultaneous reliability, n = number of items in the scale, σ_i = SD of each item, and σ_T = SD of the whole scale. Acceptable values for simultaneous reliability were considered to be 0.7–0.9.[11]

Test–Retest Reliability

The point estimate of the reliability of the NACS, its subscales, and the group of babies in the same state for both examinations was determined by calculating the ICC between observers using a fixed-effects model and the following equation:

$$R = \left(\frac{\sigma_{patients}^2}{\sigma_{patients}^2 + \sigma_{error}^2}\right) \quad (2)$$

where R = the ICC of interest, $\sigma_{patients}^2$ = the variance

Table 1. Demographic Data

Variable	Mean (SD) or n/N
Gestational age (weeks)	39.3 (1.0)
Weight (g)	3,508 (400)
Maternal age (yr)	32.2 (5.2)
Nulliparous/multiparous	65/135

of the scores among patients, and σ^2_{error} = the variance of scores between observers. Acceptable test–retest reliability was considered to be an ICC greater than 0.6.

We computed the endorsement of each item on the scale as a further description of the NACS. An item was endorsed if it received a perfect score of two. Items in the scale that were considered useful for differentiating between subjects had an endorsement rate of less than 85% (*i.e.*, < 85% of the responses equal 2).

Sample Size

The confidence interval (CI) is inversely related to the ICC. We assumed that the NACS would have an ICC of 0.6 or greater and the 95% CI would be ± 0.1. Therefore, 200 patients would be needed. If the reliability were 0.8 or greater, the CI would be ± 0.05 with the same number of patients.[11]

Results

Two hundred twenty-two babies were examined, of which 22 were excluded for the following reasons: four parents refused the second examination, seven babies were transferred to the postpartum ward between examinations, two babies underwent blood work between examinations, one baby had respiratory distress after the first examination, and eight babies did not have a second examiner available.

Demographics

Table 1 shows the maternal and newborn demographics. Fifty-seven neonates were delivered by cesarean section. There were 143 vaginal deliveries, of which 121 were spontaneous. The number of infants exposed to each drug (sedative, antiemetic, opioid, or local anes-

Table 2. Maternal Intrapartum Medications

Drug	n*	Mean Dose	Range
Bupivacaine	144	32.4 mg	1.25–246 mg
Lidocaine	71	173 mg	23–620 mg
Ropivacaine	46	84.8 mg	8.1–316 mg
Fentanyl	156	141 μg	10–703 μg
Morphine	50	0.194 mg	0.15–3.0 mg
Meperidine	5	100 mg	100–100 mg
Sufentanil	40	22.4 μg	5–240 μg
Dimenhydrinate	7	42.9 mg	25–50 mg
None	12		

* Some infants were exposed to more than one drug.

Table 3. Distribution of NACS

	Observer 1	Observer 2
Total NACS, maximum score = 40, all infants (N = 200)		
Mean (SD)	33 (3.1)	33 (2.8)
Median (range)	33 (20–40)	33 (22–38)
Total NACS, maximum score = 40, infants in the same state (n = 84)		
Mean (SD)	33.8 (3.0)	32.7 (3.2)
Median (range)	34 (20–40)	33 (22–38)
Adaptive capacity, maximum score = 10, all infants (N = 200)		
Mean (SD)	7.54 (1.9)	7.18 (1.8)
Median (range)	8 (1–10)	7 (0–10)
Neurologic assessment, maximum score = 30, all infants (N = 200)		
Mean (SD)	26.0 (2.5)	25.5 (2.7)
Median (range)	26 (13–30)	26 (16–30)

NACS = Neurologic and Adaptive Capacity Score.

thetic) is shown in table 2. Local anesthetics as a group were the most frequently used, and fentanyl was the most common opioid. One woman had a general anesthetic, 12 had an unmedicated vaginal delivery, and 4 received only nitrous oxide.

Eight infants had a 1-min Apgar score less than 7. All 5-min Apgar scores were greater than or equal to 7. Fifty infants required some resuscitative intervention. Suction was required by 47 infants, 31 received oxygen by mask, 5 received mask continuous positive airway pressure, and 5 required positive pressure ventilation. Some infants had more than one intervention.

The distribution of the NACS is shown in table 3. Two NACS observations on each of 200 infants were analyzed. Of these, 84 were in the same state at the beginning of each examination. Eight of the 20 items were endorsed in more than 85% of scores.

Simultaneous Reliability

The α coefficient was 0.47 for the total NACS, 0.42 for the adaptive component, and 0.48 for the neurologic component.

Test–Retest Reliability

Figure 1 shows the relation of the total NACS between observers. The ICC was 0.38 (95% CI, 0.24–0.52) for all pairs of observations and 0.37 (95% CI, 0.15–0.59) for babies in the same state. At hospital 1, the ICC was 0.25 (95% CI, 0.05–0.45), and at hospital 2 it was 0.42 (95% CI, 0.22–0.62). The ICC for the adaptive capacity subscale was 0.21 (95% CI, 0.07–0.34), and for the neurologic examination subscale it was 0.51 (95% CI, 0.38–0.64).
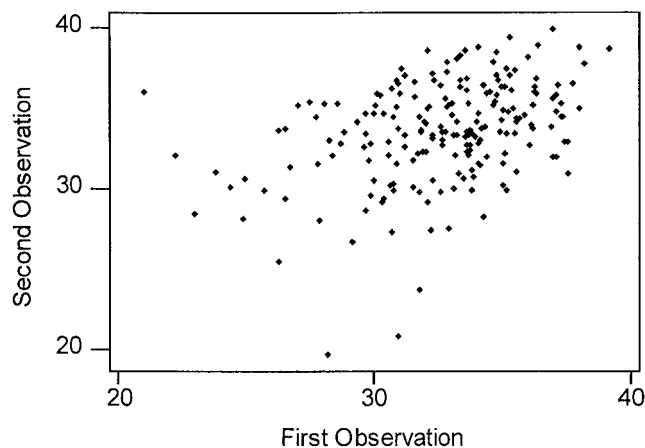
Fig. 1. Dot plot of the Neurologic and Adaptive Capacity Score. Each dot represents one patient. The x-axis represents the first score, and the y-axis represents the second score.

Discussion

There are two important aspects of reliability that must be addressed before a measurement tool can be considered for use. The first evaluates the ability of independent observers to agree on what behaviors are occurring,[12] and the second concerns stability of the measurement over time (test–retest reliability).

Data on observer agreement is presented in the original NACS report.[1] This type of reliability has the advantage of simultaneous assessment, which eliminates the possibility of the infants' behavior changing over time. However, the observers are not "independent" because the test is performed by one observer and scored by the others. An important property of reliability, stability over time, cannot be assessed. In addition, there are a number of items that depend on the observer's ability to elicit a particular response or to feel the tone of the infant. If one observer performs the test and the others simply watch, they are limited in their ability to detect subtle changes in tone.

In our study, we assessed both the simultaneous reliability of the NACS and its stability over time (test–retest reliability). Simultaneous reliability can be accomplished by arbitrarily dividing the scale into two parts and correlating the scores of each half (split-half reliability).[10] Because there are many ways to "divide" the scale into two halves, we computed the α coefficient, which gives a measure of the average of all split-half combinations. When compared with allowing one observer to perform the examination and others to watch, this approach has the advantage of allowing each observer to elicit all of the items in the scale. This measure represents the average correlation of all possible ways of splitting the NACS into two components and is equivalent to effectively administering two tests at once and assessing the correlation.[10]

We used test–retest reliability to demonstrate that the NACS was reproducible and stable over time. Factors that may compromise reliability were controlled.[10] For example, we used an extensive training period for the observers and standardized the items. We attempted to eliminate subjective interpretation by achieving consensus on the technique of examination and ensuring that each observer understood the scoring system. We controlled the environment for sound, light, and temperature as much as possible and did not retest babies if they had undergone an important change in location or an invasive examination in between the tests (*e.g.*, blood sampling). The time between test periods was chosen to be long enough to avoid fatigue in the neonate but short enough to minimize the change in scores that might occur. Many of the infants were exposed to a number of different drugs *in utero*, with bupivacaine and fentanyl being the most common. Both of these drugs have a long terminal half-life (approximately 4–6 h);[13,14] therefore, the levels would not be expected to change during the test period.

The NACS has proven to have poor reliability both on simultaneous testing and in the test–retest setting. This may be a result of several factors. The first factor may be that the instrument is flawed. Although the items were chosen carefully by the investigators, in our study, 40% were endorsed more than 85% of the time, making them ineffective in differentiating infants with high scores from those with low scores. This observation calls into question the sampling technique for the items. In addition, the use of a three-point scale may not give the examiner enough choice to adequately describe the infant. A second factor may be that the infant changed between observations. Even when the analysis of the data was confined to infants that appeared to be in the same state for both evaluations, there was poor agreement of scores. Nevertheless, the second test may differ simply because the NACS allows for items to be repeated and performed in a different sequence to obtain the "best" score. An infant that becomes mildly irritable during the examination may lose a point for consolability but gain several points for better tone. Thus, like the Brazelton neonatal assessment, there appears to be no optimal score, and statistical manipulations that depend on equal interval scaling and independence of items may not be appropriate.[3]

In summary, it is important to have a test that can be used to reliably detect the effects of intrapartum drugs and other interventions on the neonate. We demonstrated that the NACS has poor reliability both on simultaneous testing and in the test–retest situation when used for this purpose. If this goal is to be achieved, a sophisticated, functional tool needs to be developed. New scales should be rigorously assessed for reliability and validity before they are used as a research tool or in clinical practice.

References

1. Amiel-Tison C, Barrier G, Shnider SM, Levinson G, Hughes SC, Stefani SJ: A new neurologic and adaptive capacity scoring system for evaluating obstetric medications in full-term newborns. ANESTHESIOLOGY 1982; 56:340–50

2. Scanlon JW, Brown WU Jr, Weiss JB, Alper MH: Neurobehavioral responses of newborn infants after maternal epidural anesthesia. ANESTHESIOLOGY 1974; 40:121–8

3. Als H, Tronick E, Lester BM, Brazelton TB: The Brazelton Neonatal Behavioral Assessment Scale (BNBAS). J Abnorm Child Psychol 1977; 5:215–31

4. Tronick E: A critique of the neonatal Neurologic and Adaptive Capacity Score (NACS). ANESTHESIOLOGY 1982; 56:338–9

5. Michenfelder JD: Accept, revise, reject or punt: An example of the latter (editorial). ANESTHESIOLOGY 1982; 56:337

6. Brockhurst NJ, Littleford JA, Halpern SH: The Neurologic and Adaptive Capacity Score: A systematic review of its use in obstetric anesthesia research. ANESTHESIOLOGY 2000; 92:237–46

7. Zuckerman B, Bauchner H, Parker S, Cabral H: Maternal depressive symptoms during pregnancy, and newborn irritability. J Dev Behav Pediatr 1990; 11:190–4

8. Peter EA, Wilson DJ, Douglas MJ, Janssen P: Inter-observer variability associated with neonatal assessment using the Neurological and Adaptive Capacity Score (NACS) (abstract). ANESTHESIOLOGY 1999; 89:A22

9. Amiel-Tison C: A method for neurologic evaluation within the first year of life. Curr Probl Pediatr 1976; 7:1–50

10. Nunnally JC: Introduction to Psychological Measurement. New York, McGraw Hill, 1970, pp 107–31

11. Streiner DL, Norman, GR: Health Measurement Scales: A Practical Guide to Their Development and Use, Second Edition. Oxford, Oxford Medical Publications, 1995, pp 104–27

12. Mitchell SK: Interobserver agreement, reliability, and generalizability of data collected in observational studies. Psychol Bull 1979; 86:376–90

13. Magno R, Berlin A, Karlsson K, Kjellmer I: Anesthesia for cesarean section IV: Placental transfer and neonatal elimination of bupivacaine following epidural analgesia for elective cesarean section. Acta Anaesthesiol Scand 1976; 20:141–6

14. Koehntop DE, Rodman JH, Brundage DM, Hegland MG, Buckley JJ: Pharmacokinetics of fentanyl in neonates. Anesth Analg 1986; 65:227–32