Sleep Medicine 90 (2022) 167–175

Contents lists available at ScienceDirect

Sleep Medicine

journal homepage: www.elsevier.com/locate/sleep

Original Article

Creating an optimal observational sleep stage classification system for very and extremely preterm infants

E.R. de Groot [a, *], A. Bik [a, 1], C. Sam [a, 1], X. Wang [a], R.A. Shellhaas [b], T. Austin [c], M.L. Tataranno [a], M.J.N.L. Benders [a], A. van den Hoogen [a, d], J. Dudink [a, e]

[a] *Department of Neonatology, University Medical Center Utrecht, Wilhelmina Children's Hospital, Utrecht, the Netherlands*
[b] *Department of Pediatrics, University of Michigan, Ann Arbor, MI, USA*
[c] *Department of Paediatrics, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK*
[d] *Princess Maxima Center for Pediatric Oncology, Utrecht, the Netherlands*
[e] *Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, the Netherlands*

ARTICLE INFO

Article history:
Received 2 September 2021
Received in revised form
18 January 2022
Accepted 22 January 2022
Available online 1 February 2022

Keywords:
Preterm infants
Sleep-wake stages
Behavioral observation
Neonatal intensive care

ABSTRACT

Background: Sleep plays a major role in neuronal survival and guiding the fetal brain's development. Preterm infants in the neonatal intensive care unit are exposed to numerous external stimuli that can severely disrupt their sleep/wake patterns. Currently, almost no behavioral classification scales are validated for preterm infants. This study aims to develop a new, easy-to-use, validated visual sleep stage classification system for preterm infants with a gestational age between 25 and 37 weeks.

Methods: The Behavioral Sleep stage classification for Preterm Infants (BeSSPI) consists of four sleep-wake stages; active sleep (AS), quiet sleep (QS), intermediate sleep (IS) and wake (W), which are classified using seven items. Items include eye movements, body movements, facial movements, vocalizations, heart rate, respiratory pattern and activity level.

Results: 69 preterm infants were observed (24 + 6–36 + 0 weeks GA at birth; 25 + 2–36 + 6 weeks PMA at observation; 57.3% male). Across all 69 infants, the BeSSPI was based on 10,922 min of observed behavior, with 4264 min AS (38.83%), 2873 min QS (26.16%), 2887 min IS (26.29%), and 957 min W (8.72%). For the final BeSSPI, an interrater agreement of $\kappa = 0.80$ was reached. Additionally, construct, content, face validity, and expert validity were carefully assessed and deemed satisfactory.

Conclusions: We developed a method to evaluate sleep-wake stages that is simple for all neonatal healthcare providers to learn and use. The BeSSPI is of high reliability and validity. Furthermore, it can be used in all preterm age-groups. Therefore, this novel instrument may improve rigor and reproducibility for future preterm sleep research.

1. Introduction

Worldwide, approximately 15 million children are born prematurely each year [1]. The sudden change in environment from the safety of the womb to an incubator in the neonatal intensive care unit (NICU) affects several key developmental milestones, often with long-lasting consequences [2–9]. Additionally, up to half of all infants born preterm are at risk of developing lifelong health complications, including neurodevelopmental and behavioral disabilities [10–12].

During the neonatal intensive care period, several factors such as inflammation, hypoxia, stress, and medication, can have negative impacts on preterm brain development [13–18]. However, little is known about the exact role of preterm sleep and wake stages on brain development. This knowledge gap is particularly surprising because behavioral states and neuronal circuits develop simultaneously. However, it is unclear if sleep and brain development are merely temporally related and otherwise independent, or if one drives the other – and if so, which way around [19]. Moreover, it seems that sleep–wake states differ significantly between a fetus in the third trimester and a preterm infant in the NICU [20].

* Corresponding author. Department of Neonatology, Wilhelmina Children's Hospital, University Medical Center Utrecht, Lundlaan 6, 3584 EA, Utrecht, the Netherlands.
E-mail address: e.r.degroot-31@umcutrecht.nl (E.R. de Groot).

[1] Authors contributed equally.

https://doi.org/10.1016/j.sleep.2022.01.020
1389-9457/

Abbreviations

AS	Active sleep
BeSSPI	Behavioral sleep stage classification for preterm infants
BSSC	Behavioral sleep stage classification
CA	Conceptional age
GA	Gestational age
HR	Heart rate
IS	Intermediate sleep
nCPAP	Nasal continuous positive airway pressure
NICU	Neonatal intensive care unit
nIPPV	Non-invasive positive pressure ventilation
PDCA	Plan-do-check-act
PMA	Postmenstrual age
PSG	Polysomnography
QS	Quiet sleep
REMs	Rapid eye movements
RP	Respiratory pattern
S-IMV	Synchronized intermittent mandatory ventilation
sIPPV	Synchronized intermittent positive pressure ventilation
W	Wake

To obtain knowledge on the effects of preterm sleep, information on sleep stage distribution, duration and quality must be effectively and efficiently monitored. In addition, preterm sleep quality can be used as a diagnostic and predictive tool for neuronal health and as a real-time monitoring tool for overall wellbeing [21–23]. Finally, the ability to monitor sleep in real-time may allow medical staff to plan elective caregiving with minimal disruption of the sleep cycle [24].

To our knowledge, currently in literature no consensus exists on how to assess sleep in preterm born infants. This makes it increasingly difficult to interpret results between studies. Polysomnography (PSG) and behavioral observations are used most frequently. However, PSG requires attachment of additional sensors and electrodes, which can cause skin damage and consequently increase risk for infection [25], require significant expertise for interpretation, and results are not available in real time. Finally, EEG measures may also be used, however they seem to be less reliable in preterm infants <28 weeks postmenstrual age (PMA) [26]. Two well-known polygraphic scales have been developed by Curzi-Dascalova and Mirmiran [27] and Anders, Emde and Parmelee [28]. Both scales provide extensive information on using a combination of EEG measures, vital parameters and behavioral observations to classify sleep stages. Interestingly, the observational criteria described by Curzi-Dascalova [27] and Anders [28] are frequently used separately to classify sleep stages [29]. However, no published evidence exists on validity of using the observational part of these scales as a stand-alone method for preterm infants.

Nevertheless, the value of behavior observational sleep stage classification (BSSC) is undeniable. BSSC seems to be one of the most versatile ways to monitor sleep stages in preterm infants. Behavioral observations are non-obtrusive, possible in all settings and even used in extremely preterm infants [29]. One of the most used standardized BSSCs is the manual by Brazelton and colleagues [30]. In this manual, the active sleep stage (AS) was defined by closed eyes, rapid eye movements (REMs), low activity levels, random movements, startles, responsiveness to internal and external stimuli, irregular respiration, and sucking. Quiet sleep stage (QS) was defined by closed eyes, no eye movements, no spontaneous activity, occasional startles and jerks, and regular respirations. Several other methods have been published based on similar criteria [31–33]. Other frequently used and well-known methods were developed by Prechtl [34], Stefanski [35], and Thoman [36].

It is important to note that of the previously mentioned BSSCs, only Stefanski's method was developed using a preterm sample. Thus, to date, preterm sleep has mainly been studied using methods originally developed for full-term infants. Yet, full-term infants show distinct sleep architecture [37], making the reliability of these methods in preterm samples questionable. Furthermore, despite rapid progress in neonatal sleep research, sleep studies in extremely preterm infants (<28 weeks postmenstrual age; PMA) and very preterm infants (<32 weeks PMA) have remained relatively underemphasized. Until now, only two BSSCs have been developed for extremely preterm infants [38,39]. However, their reliability and validity have not been assessed. In short, a fully valid and reliable BSSC to study extremely preterm infants does not currently exist, although sleep seems to be one of the most important protective factors for proper brain development [40].

The current study aims to develop an optimal system to classify sleep stages in preterm infants. Such 'optimality' will comprise a new, easy-to-use, validated and reliable behavioral sleep stage classification system that can be applied to extreme and very preterm infants, within the first postnatal days and weeks. We aim to develop a Behavioral Sleep stage classification for Preterm Infants (BeSSPI) that is easily reproducible and could be easily used in most clinical settings without additional technical requirements and knowledge. Furthermore, it is essential that sleep stage classifiers be readily implemented in clinical care worldwide to guide studies and quality improvement work designed to promote sleep in preterm infants.

2. Methodology

The Behavioral Sleep stage classification for Preterm Infants (BeSSPI) was developed using a plan-do-check-act (PDCA) cycle (see Supplement A for an overview). Sleep stage definitions were based on two systematic literature reviews conducted by our group. The first considered heart rate and respiratory frequency in preterm sleep [41], and the second encompassed an extensive overview of previously developed BSSCs [29]. The second review also elaborated on the reliability and validity of BSSCs and the most-used scoring items. The items that emerged from this review were used as the basis for the BeSSPI.

During the first PDCA cycle the first version of the scale was developed and the interrater agreement (Fleiss' kappa) was calculated to assess reliability of the scale ('plan' phase), using video data. Afterwards, observers independently classified a group of preterm infants to gain experience with preterm sleep behavior ('do' phase). At the end of each PDCA cycle, the interrater agreement was discussed among the researchers and the scale was adapted based on the observers' experience during the separate observations ('check' and 'act' phase). Then the scale was explained to a new group of researchers, who repeated the process. The final research team combined notes from all PDCA cycles into a final scale. Validity and reliability of the final scale was assessed extensively (see also Supplement A: Fig. S1).

2.1. Participants and setting

A population of 69 preterm infants between 25+0-weeks and 36+6-weeks PMA (born between 24+6-weeks and 36+0-weeks GA) was included for observation to develop the final BeSSPI. The infants were all admitted to the NICU of the Wilhelmina Children's Hospital in Utrecht. The exclusion criteria in all cases were infants diagnosed with congenital malformations, seizures, and significant brain injury (eg, intraventricular hemorrhage grade 3 or 4, post-hemorrhagic ventricular dilatation, and cystic periventricular leukomalacia). Infants who were currently receiving sedatives or invasive respiratory support were also excluded, as were infants whose mothers' used narcotics during pregnancy.

Parents were informed regarding the observational research, and they gave written informed consent. The research protocol (21–066C) was presented to the Medical Research Ethics Committee (METC), who confirmed that the Medical Research Involving Human Subjects Act (WMO) does not apply to this study.

2.2. The BeSSPI

The sleep/wake cycle was divided into four stages for the BeSSPI: active sleep (AS), quiet sleep (QS), intermediate sleep (IS), and wake (W). Note that in the BeSSPI 'IS' refers to intermediate sleep – not indeterminant sleep, which is commonly recorded in other infant sleep staging systems – intermediate sleep (IS) is considered a transitory stage. Therefore, IS is not used to indicate epochs during which it was unclear if it should be classified as AS or QS (see 2.2.4. Intermediate sleep and 2.2.8. Confidence scores for more information). See Table 1 for the description of these stages. Stages were classified in 1-min epochs (see 2.2.2. Epochs below). The observations took place during the daytime in 1- to 3-h time slots. Observations were performed by two to three independent researchers per PDCA cycle (seven researchers in total; in the last PDCA cycle this were EG, AB, CS).

Observations were either performed live, next to the incubator, or using video footage when permission was given to record the infants. Videos were recorded using two synchronized camera's – one on a portable tripod recording the infant and one recording the vital signs monitor. An observation form was used to record information regarding the determined sleep characteristics (see Supplement B). Observations were not performed during kangaroo care (skin-to-skin care) and were planned before or after routine scheduled nursing care, where possible, to avert interference with the observations.

2.2.1. Items

Seven items were used to distinguish between sleep stages (for specifics see Supplement C). These items included eye movements, body movements, facial movements, vocalizations, heart rate, respiratory pattern and activity level, with eye movements being the most important. Rapid eye movements (REMs) are essential in distinguishing AS from QS; REMs do not occur during QS. Furthermore, face and body movements were carefully considered. Specific movements can occur during both AS and QS. When this is the case, activity level was used to determine the sleep stage. A high activity level, which is characteristic for AS, was defined by the following:

1. Movements occurring for the majority of two 15-s periods within a single 1-min epoch.
2. Movements occurring for more than 15 consecutive seconds.
3. More than three separate movements in an epoch. Rhythmic movement series, such as mouth/sucking movements, were counted as one movement.

With this, activity level served both as a 'summary' of the perceived activity during an epoch and as a cut-off when movements that can occur during QS occur in high frequency. In addition, vocalizations were assessed, however, vocalizations were difficult to distinguish when observing an infant in a closed incubator.

Finally, heart rate (HR) and respiratory pattern (RP) were measured using the 'Philips Intellivue MP70 Neonatal monitors' bedside monitor (Koninklijke Philips N.V., Eindhoven, The Netherlands). We took into account that age influences cardiorespiratory parameters in the following ways [41]:

2.2.1.1. Respiratory pattern

- Increased postnatal age is related to generally faster RP [41].

- There is also an indication that:
  o RP is higher in QS than in AS for 27–32 wk PMA [41].
  o RP does not differ between QS and AS for 31–34 wk PMA [41].
  o RP is higher in AS than QS for late preterm infants (>35 wk PMA) [41].

Table 1
Sleep stage definitions. A general overview of characteristics that can occur during different stages. Note that in the current system intermediate sleep (IS) is considered a transition stage, which is only classified when a clear transition between AS and QS or between wake and sleep occurs (for more information, see '2.2.4. Intermediate sleep'). Item characteristics are based on the in-house developed BeSSPI. For an extensive overview and explanations, see Supplement C.

	Active Sleep	Quiet Sleep	Wake
Eyes	Closed Rapid eye movements (closed or slightly opened eyes)	Closed	Open
Body movements	Gross movements Small movements	Reflexive movements High muscle tension[a] No movements	Gross movements No movements
Facial movements	Non-reflexive facial movements Non-rhythmic mouth movements	Reflexive facial movements Rhythmic mouth movements No facial movements	Full range of facial movements
Vocalizations	Grunts Distressed sounds Reflexive sounds	Sobs/Sighs Reflexive sounds	Full range of vocalizations
Heart rate[b]	Irregular	Regular	Regular, but faster
Respiratory pattern[b]	Irregular	Regular	Regular, but faster
Activity level	High	Low	Either high or low

[a] High muscle tension was assessed by visual observation [29]. Writhing, stirring or stretching may also be considered high muscle tone, however these characteristics are mainly considered an AS characteristic [29].
[b] Heart rate and respiratory pattern assessment was based on observation of the bedside monitor.

2.2.1.2. Heart rate

- Increased gestational age (GA) at birth is related to lower HR [41].

- Increased postnatal age is related to higher HR [41].

- Increased GA at birth is related to higher heart rate variability (HRV; mainly in AS) [41].

- Increased postnatal age is related to higher HRV [41].

2.2.1.3. Cardiorespiratory coupling

- Cardiorespiratory synchronization/coupling only starts to occur between 27- and 32-weeks PMA [41]. So, before 27–32 weeks PMA, HR and RP cannot be expected to show similar patterns. If there was a discrepancy between HR and RP, other items were considered with more weight.

When an infant was of lower postnatal age, the observers took into account that

1) between 31 and 34 weeks PMA, RP may differ less between QS and AS;
2) before 33 weeks PMA, HR be generally lower; and
3) before 32 weeks PMA irregular HR may not always be accompanied by irregular RP (and vice versa).

Finally, HR and RP may become more irregular during startles, stretches, jerks, apneas, and bradycardias. When such a reflexive body movement or cardiorespiratory event occurs, an irregularity in HR/RP not necessarily indicates that the epoch should be classified as AS.

For a summarized overview of all items and visual aid to be used during observations, see Fig. 1.

2.2.2. Epochs

Epochs were 1 min, which we found to be long enough to observe all parameters multiple times and short enough that sleep stages were unlikely to vary during the epoch. It is important to mention that this deviates from polysomnograms, which are divided into 30 s epochs. However, we found 30 s was too short to reliably register all characteristics multiple times. To prevent the influence of one overlooked behavior to bias the whole epoch, we decided to elongate the epochs. Additionally, the observations were

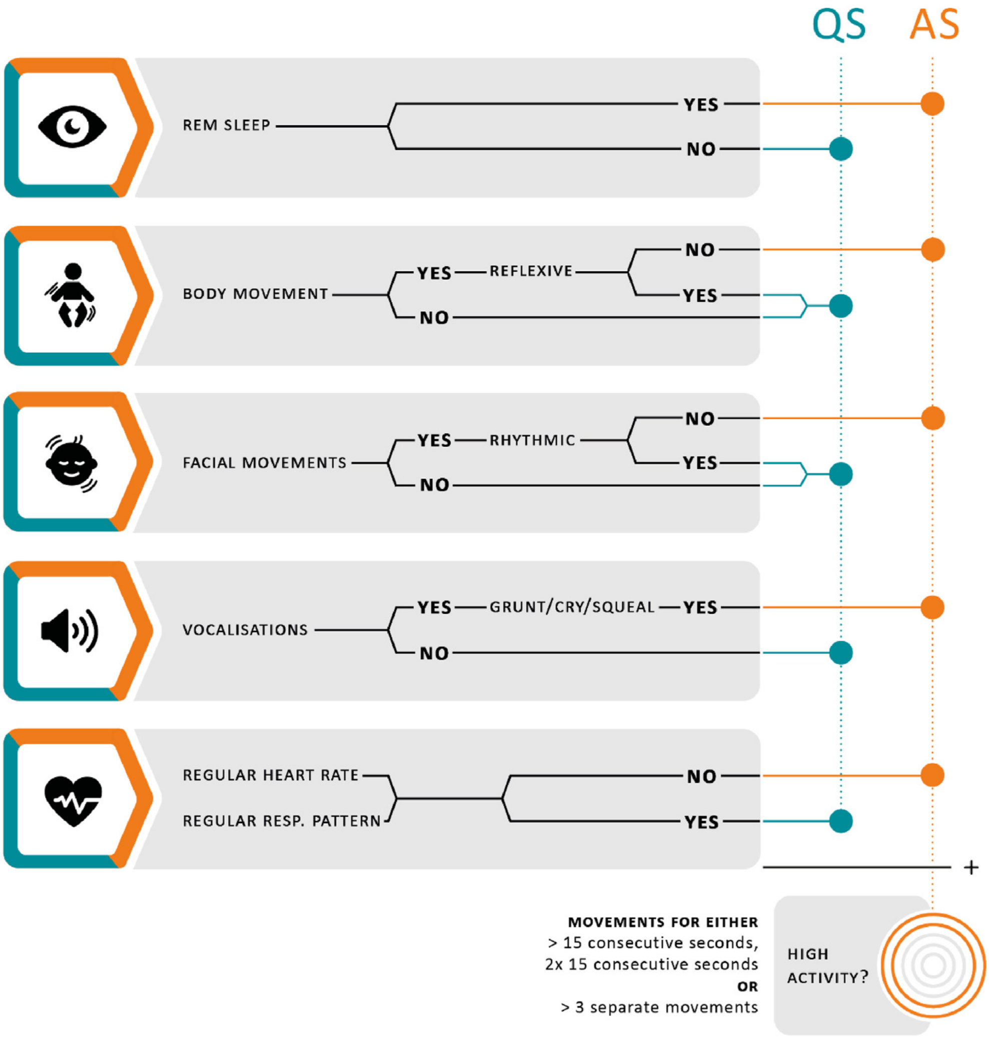

Fig. 1. A schematic overview of the AS vs. QS classification of the BeSSPI. In the flowchart only distinctive characteristics between quiet sleep (QS) and active sleep (AS) are displayed.

divided into 15-s intervals within these epochs, to assess each parameter at least four times.

2.2.3. Baseline observation

As some infants are generally more active than others or have more variable HR and RP baselines, a 15-min baseline observation was undertaken to adapt the observation to these individual differences. During these baseline observations synchrony between HR and RP was taken into account, as well as the amount of face and body movements. This knowledge is then used to further inform choices with regard to sleep stage. For example, if relatively regular HR and RP are seen in combination with a relative high number of twitches and the absence of REMs, the observer knows that the baseline body-activity may be high. Behavior was classified in sleep stages using the BeSSPI, but these classifications are not used in further analyses.

2.2.4. Intermediate sleep

In the BeSSPI, intermediate sleep (IS) was considered a transition stage between QS and AS, or between sleep and W. During observations, it can be difficult to see transitions in real-time due to the disorganized nature of extremely preterm sleep. Therefore, the focus was mainly on QS and AS classifications. An epoch was classified as IS only when there was a clear transition stage.

2.2.5. Drowsy

Infants enter a stage that the literature frequently calls “drowsy” when transitioning from wake to sleep. We classified this stage as “IS”. The drowsy stage is frequently characterized by open, unfocused eyes (“glassy eyes”) or opening and closing the eyes and may strongly resemble AS with REMs. However, there are some characteristics of an infant being drowsy that do not frequently occur during AS:

1. Variable activity levels with relatively little body movement [30,36,42].
2. Vocalizations (sometimes fussy) [42,43];
3. Regular respiratory pattern, but higher than during QS [43–45];
4. Possible reactivity to sensory stimuli [30];
5. Often occurs after wake.

2.2.6. Wake

While the infant is awake, different types of behavior can be distinguished; the infant may be calm, fussy, or highly responsive and the eyes are open. However, these stages were not distinguished in the BeSSPI, as it focusses on sleep behavior.

2.2.7. Crying

When an infant was crying (C), this was generally considered as W. However, sometimes crying could occur during sleep (AS or IS) in the following manner:

1. The infant cries for less than 5 s.
2. Crying starts and stops instantly.
3. The behavior before and after crying clearly resembles sleep.

In this case, the epoch was classified as AS if the sleep behaviors before and after crying represent AS. The epoch was classified as IS if the sleep behaviors before and after crying represent QS or IS.

2.2.8. Confidence scores

Confidence scores were added after an epoch was classified to note the observer's confidence in the classification, using the following scores:

1 = Highly confident that this sleep stage is correct (80–100% confidence).

0 = Relatively confident that this sleep stage is correct (50–80% confidence).

−1 = Not confident that this sleep stage is correct (0–50% confidence).

A confidence score of 0 or -1 is different from classifying a stage as IS: if a stage is clearly disorganized or resembles a transition, an IS stage was recorded. If the observer thought they may have misinterpreted an item that is non-characteristic for the current stage, such as REMs during QS that may have been an eye-twitch, they were able to decrease the confidence score and retrospectively consider the more plausible event, a transitory stage or just a misinterpreted twitch.

2.2.9. Smoothing

As the BeSSPI was relatively complex and the researchers may not have classified each epoch with 100% confidence, reclassification and smoothing were applied retrospectively. With smoothing, epochs with lower confidence scores (0 or −1) were reassessed based on the information recorded on the observation form. Reassessment was be done by using the manual, discussing with well-trained fellow researchers, or consulting the literature.

Given the heterogeneous nature of infant sleep states, particularly relating to age and sleeping position [41], age, position, and interventions were recorded to allow these characteristics to be applied in retrospective smoothing.

2.3. Data analysis

Interrater reliability (Fleiss' kappa; [46]) was calculated during each PDCA cycle. Interrater reliability measures the degree of agreement between raters, corrected for the extent of agreement expected by chance alone. Fleiss' kappa is specifically developed for measuring the agreement among any (constant) number of raters, as opposed to Cohen's kappa, which is only fit for calculating agreement between two raters at a time [46]. All interrater reliability scores are calculated with video data. In this way it was certain that all observers could see the same information.

In addition to the observers' agreement on sleep stage classification as a whole, agreement on different characteristics, such as occurrence of REMs, was calculated to determine if disagreement between observers was caused by differences in identification or interpretation. Finally, the confidence score helped to determine if disagreements also occurred when observers were highly convinced of their classification.

Furthermore, content and construct validity were assessed during the development of the BeSSPI to determine to what degree items the individual represented the measured construct (ie sleep stage) and whether the items covered the full range of the construct [47]. First, an over-complete version of the BeSSPI was created, including all items identified in the literature and all behaviors that we frequently observed in our clinical practice. Subsequently, the literature was used to identify which items were similar and could be rationally combined [29].

Finally, to determine expert and face validity, qualified clinicians with extensive experience in classifying preterm sleep (Dr. Jeroen Dudink and Dr. Renée Shellhaas) were consulted for their feedback on the BeSSPI. Both clinicians systematically assessed the method used to develop the BeSSPI and the items used to classify sleep.

3. Results

Eight observations were done using 5 videos and 69 observations were done at the bedside. All video observations were used to calculate interrater reliability (see below). We made sure that only videos were obtained from the 69 infants that were observed at the bedside, within a few days of bedside observation. For bedside observations the median age of infants was 30 weeks and 2 days PMA at observation, ranging from 25 weeks to 2 days to 36 weeks and 6 days. The median age of infants at birth was 28 weeks and 4 days GA, ranging from 24 weeks to 6 days to 36 weeks. Patient characteristics are shown in Table 2.

3.1. Inter- and intra-rater agreement

For the first three PDCA cycles, reliability showed a decreasing trend, with a Fleiss' kappa of κ = 0.76, κ = 0.48 and κ = 0.46, respectively. Each PDCA cycle was completed by a different research group, consisting of research interns with a medical, neuroscience, biology and/or psychology background (Supplement A). For a complete overview of all interrater agreements, see Supplement D.

An interrater agreement of κ = 0.80 for all sleep stages was reached for the final version of the BeSSPI, both during the first observation and the retest. Furthermore, intrarater agreements of κ = 0.93–0.79 were reached. Interrater and intrarater agreements ranged between κ = 0.41–1.00 for individual items (see Table 3). After removing the first 15 min of the observation (ie the baseline), interrater agreements increased by 0.01–0.15 points. All interrater agreements were calculated between three observers (EG, AB and CS).

3.2. Validity

Construct and content validity were continuously assessed. All known aspects of sleep behavior were considered by basing the BeSSPI on previous behavioral observation methods. All items included in previous methods or that we saw ourselves during observation, such as the occurrence of hiccups inducing stage change, were included in the final BeSSPI. After thoroughly evaluating the usefulness of the categories as a whole, redundant items were combined or removed from the BeSSPI.

Furthermore, the BeSSPI was checked for expert and face validity by two healthcare professionals with expertise in preterm sleep-behavior (JD and RS). The cyclic method of adapting the scale and calculating multiple interrater agreements was considered to be a precise and comprehensive way to develop an observational scale. Items corresponded with the expected behaviors, based on clinical and scientific expertise.

3.3. Data gathered using the new classification

The BeSSPI was based on 10,922 min of observed behavior (69 infants), with 4231 min AS (38.74%), 2868 min QS (26.26%), 2866 min IS (26.24%), and 957 min W (8.76%) (see Table 4).

During the 10,922 min of observation, 191 interruptions took place (see Table 5). These interruptions all involved essential handlings by a caregiver (nurse, parent, or doctor) opening the incubator. Interruptions included but were not limited to feeding, checking lines, care after bradycardia or apneas, and ultrasounds. If possible, observations were continued during interruptions. However, if visibility of the infant decreased too much, the epoch was classified as not available (NA).

4. Discussion

We have developed and validated an easy-to-use. new BSSC (the BeSSPI) for preterm infants from 25 to 37 weeks PMA, within the first postnatal days and weeks of life. The BeSSPI had very good inter- and intra-rater reliability and should be readily applicable for clinical, research, and quality improvement purposes in the NICU. The process included four PDCA cycles, conducted over three years. This extensive process resulted in the final BeSSPI and included an in-depth overview of expected behaviors during different sleep stages. The BeSSPI achieved a high interrater agreement of κ = 0.80.

Compared to previous scoring systems [29], the BeSSPI is unique in several ways: first, the influence of age is considered when assessing heart rate and respiratory pattern characteristics; second, the BeSSPI includes confidence scores to be indicated per epoch by observers; third a section on "smoothing" the data was added, allowing classifications to be adapted retrospectively to increase accuracy; and fourth, a decision flowchart was created to improve usability (Fig. 1).

Table 2
General patient characteristics of infants per scale.

	PDCA cycle#1	PDCA cycle#2	PDCA cycle#3	Total
Number of infants	N = 36	N = 18	N = 15	N = 69
Sex	M = 19	M = 11	M = 8	M = 38
	F = 17	F = 7	F = 7	F = 31
GA at birth	29w + 2d	26w + 4d	27w + 5d	28w + 4d
Median (range)	(25 + 2–36 + 0)	(24 + 6–34 + 3)	(25 + 1–31 + 1)	(24 + 6–36 + 0)
Birth weight	1274 g	1003 g	1135 g	1173 g
Median Apgar score (1/5/10 min)	6/8/8	6/7/8	5/7/7	6/8/8
IQR Apgar score (1/5/10 min)	5/3/1	2/1/1	2/2/1	3/3/1
PMA at observation	30w + 4d	31w	30w + 1d	30w +2d
Median (range)	(26 + 0–36 + 6)	(25 + 2–33 + 2)	(25 + 4–32 + 5)	(25 + 2–36 + 6)
Respiratory support during observation	No = 10	No = 2	No = 5	No = 17
	Optiflow = 5	Optiflow = 4	Optiflow = 0	Optiflow = 9
	nCPAP = 14	nCPAP = 6	nCPAP = 6	nCPAP = 26
	nIPPV = 6	nIPPV = 4	nIPPV = 3	nIPPV = 13
	sIPPV = 1	sIPPV/S-IMV = 2	sIPPV = 1	sIPPV/S-IMV = 4
Phototherapy during observation	N = 13	N = 0	N = 2	N = 15
Sleeping position during observation	Supine = 10	Supine = 11	Supine = 6	Supine = 27
	Lateral = 24	Lateral = 12	Lateral = 9	Lateral = 45
	Prone = 10	Prone = 4	Prone = 1	Prone = 15
Minutes observed	Total = 6017	Total = 2386	Total = 2519	Total = 10,922
	Mean per observation = 167	Mean per observation = 133	Mean per observation = 168	Mean per observation = 158

D: Days; GA: Gestational age; nCPAP: Nasal continuous positive airway pressure; nIPPV: Non-invasive positive pressure ventilation; PMA: Postmenstrual age; SD: Standard deviation; S-IMV: Synchronized intermittent mandatory ventilation; sIPPV: Synchronized intermittent positive pressure ventilation; W: Weeks.

Table 3
Overview of all interrater and intrarater agreements for the final BeSSPI.

Interrater agreement observation 1 (31 weeks PMA)		κ = 0.80	
Interrater agreement observation 2 (31 weeks PMA – retest)		κ = 0.80	
		Observation 1	**Observation 2**
Interrater agreement REMs		κ = 0.56	κ = 0.66
Interrater agreement HR		κ = 0.41	κ = 0.72
Interrater agreement RF		κ = 0.62	κ = 0.82
Interrater agreement activity level		κ = 0.59	κ = 0.61
	AB	CS	EG
Intra-rater agreement sleep stage	κ = 0.93	κ = 0.88	κ = 0.79
Intra-rater agreement REMs	κ = 0.69	κ = 0.59	κ = 0.64
Intra-rater agreement HR	κ = 0.58	κ = 0.77	κ = 0.54
Intra-rater agreement RF	κ = 0.80	κ = 1.00	κ = 0.65
Intra-rater agreement activity level	κ = 0.63	κ = 0.63	κ = 0.84

Table 4
Overview of the length of the total observation and separate sleep stages. Missing data were excluded.

	Min observation	Min AS	Min QS	Min IS	Min W
Total duration all observation (minutes)	10,922	4231	2868	2866	957
Mean duration per observation (minutes)	158.3	61.3	41.6	41.5	13.9
Min duration per observation (minutes)	48	3	0	8	0
Max duration observation (minutes)	180	123	98	79	82

Table 5
Overview of the number and length of interruptions occurring during observations.

	Mean	Median	Range
Number of interruptions per observation	2.8	3	0–9
Length of all interruptions during one observation period	10.2 min	7 min	0–38 min
Length of one interruption	3.6 min	2.3 min	0–37 min

The most important considerations when developing the BeSSPI were the sleep stages and the items used to classify them. Between 1975 and 1995, preterm sleep researchers voiced different opinions on the best way to categorize and classify sleep/wake cycling in preterm infants. For example, Prechtl [48] and Thoman [36] appear to hold opposite views, with Prechtl advocating for a dichotomic description of stages, using a limited number of straightforward criteria. On the contrary, Thoman advocated for detailed descriptions of state criteria, which are continuously measured. Furthermore, Prechtl only uses a limited number of sleep stages, whereas Thoman subdivided AS and QS in multiple substages (see also Supplement E).

The BeSSPI combined aspects of both Prechtl and Thoman's criteria. The combination between their respective scores is reflected in the continuously measured items, using 1-min epochs to classify sleep stages (Prechtl). These epochs are subdivided into 15-s intervals better to grasp the individual differences between infants (Thoman). Stage criteria consist of descriptive items (Thoman) with clear presence or absence differentiation (Prechtl). In addition, a new parameter indicating the infants' activity levels was added to increase the general continuity.

Furthermore, despite describing all possible behaviors (Thoman), the BeSSPI included meaningful clusters of items, resulting in straightforward (Prechtl) but detailed criteria. Sleep stages were divided into four interrelated categories. The newly added concept of “smoothing” increased the context-bound nature of the BeSSPI. Moreover, since the interrelated quality of sleep stages was considered after classification, errors made during classification did not influence subsequent epochs. Overall, the BeSSPI was both detailed and concise, significantly reducing the risk for misinterpretation of items or stages. Therefore, high reliability was reached between researchers.

4.1. Validity and reliability

Interrater reliability was assessed during each PDCA cycle and was assessed multiple times during the first cycle. Reliability showed a decreasing trend (from high to moderate) after one researcher, who had stopped making observations, was still included in the Fleiss' kappa calculation. Additionally, the time between the observations and the last discussion about the items resulted in a further decrease in interrater reliability. This finding indicated that continued practice and ongoing discussions regarding the difficulties experienced during observations are essential to maintain the reliability of the BeSSPI at a substantial level.

The interrater reliability for the second and third cycles was calculated once and showed a moderate agreement between researchers and a low agreement between the research groups in other PDCA cycles. During the second and third cycle a high number of new behaviors were added to the scale. This result suggested that, as well as appropriate training, a clear definition of each item and sleep stage is crucial to prevent misunderstandings. Furthermore, a higher number of different behaviors may be confusing and decrease reliability, which possibly explains the decrease in interrater agreement found during the second and third cycle. As a result, items were clustered and clear rules were established to prevent confusion in the future. This will improve clarity and trainability and ensure that different research groups using the BeSSPI classify the parameters similarly. In other words, clear rules will improve reliability between research groups, even if they cannot discuss the different items.

Regarding reliability measures, the dichotomic items were assessed for interrater agreement. The interrater reliability for

REMs, HR, RF, and activity levels ranged between κ = 0.41–0.82. The item-kappa's increased between the first and second observation of the same video, possibly due to practice effects (eg memory recall). However, observers waited one month between test–retest assessments. Another explanation for the increase was the research group's discussion regarding the epochs they did not agree upon after the first observation. Classification of complex items, such as HR and RF, was discussed extensively, potentially explaining the vast increase in interrater reliability for HR and RF.

Finally, test–retest reliability could not be measured for stability over time in the sleep stages of individual infants [36]. Preterm infants develop at a high rate, resulting in changes in the distribution and composition of their behavioral states. Therefore, test–retest reliability was measured in the research by observing the same video twice, resulting in relatively high intrarater reliability (κ = 0.79–0.93). Intrarater agreements for items (REMs, HR, RF, and activity level) ranged between κ = 0.54–1.00. Interestingly, the intrarater reliability for separate items differed between raters. For example, CS reached a kappa of κ = 0.77 on HR, whereas EG and AB reached κ = 0.54 and 0.58, respectively. This result indicated that different observers showed varied dexterity regarding the individual criteria of the BeSSPI. However, the high interrater and intrarater agreement for sleep stage classification indicate that the instrument as a whole is reliable, regardless of the individual dexterity on specific items.

Content, construct, face, and expert validity was determined sufficient after careful assessment. An extensive study by Hayes [49] into the frequency and relationship of items is consistent with the method used in the BeSSPI. Hayes' results indicated more quiescence (>5s in a 1-min epoch) during QS. Moreover, Hayes found more general motor activity during AS. Finally, the combination of AS characteristics, such as general motor activity, facial movements, and eye movements, was the most commonly occurring, independent of age. Although Hayes concluded that stages are slightly more disorganized in younger infants (<30 weeks conceptional age; CA), he also concluded that “the coalescence of stage-related behaviors is still an emergent process at (>30 weeks CA).”

4.2. Strengths and limitations

Compared to term infants, preterm sleep is generally thought to consists of twice as much AS as QS [39]. With 4231 min of AS and 2868 min QS, the distribution of sleep stages we found showed more QS than expected from the literature. However, the AS/QS division found when using the BeSSPI is quite similar to values reported by Thoman [50], Curzi-Dascalova [51] and Bourel-Ponchel [52]. Furthermore, when considering age-related differences in sleep stage distribution, our findings also showed similar results (See Supplement F).

Finally, behavioral sleep stage classification is often criticized as time-consuming due to the time it takes to train an expert and the time taken to observe and classify sleep stages. With the BeSSPI, we have tried to reduce the long training time significantly. The long observation time can be diminished by developing an automatic sleep scoring algorithm based on existing valid and reliable sleep stage classification methods. Preferably, the algorithm would use non-intrusive methods to maintain the advantages of observational sleep stage classification. For example, classifications performed with the BeSSPI can be time-linked to existing cardiorespiratory, EEG or video data. The BeSSPI annotations will then serve as input target data for the machine learning process which tries to predict sleep-wake states from vital parameters. Later, BeSSPI annotations may also be used for validation of the algorithm. In other words, when creating a supervised machine learning model, the vital parameters can serve as input and the BeSSPI annotations as output.

5. Conclusion

The BeSSPI is developed for the purpose of being simple to learn and easy to use. The provided decision flowchart (Fig. 1), classification form (Supplement B), and item list (Supplement C) will support inexperienced users. In addition, the BeSSPI provides the next step in observational preterm sleep stage classification, as it is created using both existing knowledge on preterm sleep behavior and our own observations. With this, an extensive overview was created of all possible behaviors that may occur during preterm sleep. Furthermore, by adding confidence scores, IS is not used as a ‘residual category’ (eg when an observer is unsure about their classification). Finally, the new category of ‘activity level’ is added, which increases general continuity of scoring and makes it easier to account for individual differences.

The BeSSPI has a high reliability and validity supporting its use in extremely preterm infants. Furthermore, because the development method is extensively described, it is possible to replicate the process or to adapt the BeSSPI to different contexts. Finally, the BeSSPI was developed with the involvement of multiple groups of researchers and includes clear rules specifically designed to increase reliability between research groups. With this, the BeSSPI meets all previously set criteria for an ‘optimal’ behavioral sleep stage classification method. Therefore, using this robust tool across studies will increase the replicability and comparability of preterm sleep research. In summary, the BeSSPI provides an ideal starting point for the ongoing surge of renewed interest in investigating the importance of sleep during preterm period.

Declaration of funding source

The authors acknowledged the financial support of EU H2020 MSCA-ITN-2018: INtegrating Functional Assessment measures for Neonatal Safeguard (INFANS), funded by the European Commission under Grant Agreement #813483.

Acknowledgements

We thank Charlotte Teunis, Marit Knoop, Simone Visser and Laura van Wijngaarden for helping us with observations and for providing insights about the BeSSPI during the development process. We thank Anne Roos de Gooijer for visualizing the BeSSPI manual (Fig. 1) and providing a design point of view during development.

Conflict of interest

Dr. Shellhaas receives research support from NIH and the Pediatric Epilepsy Research Foundation. She serves as an associate editor for *Neurology*, is a consultant for the Epilepsy Study Consortium, and receives royalties from UpToDate for authorship of topics related to neonatal seizures.

The ICMJE Uniform Disclosure Form for Potential Conflicts of Interest associated with this article can be viewed by clicking on the following link: https://doi.org/10.1016/j.sleep.2022.01.020.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.sleep.2022.01.020.

References

[1] Blencowe H, Cousens S, Oestergaard MZ, et al. National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. Lancet 2012;379(9832):2162–72. https://doi.org/10.1016/S0140-6736(12)60820-4.

[2] Aarnoudse-Moens CSH, Weisglas-Kuperus N, van Goudoever JB, et al. Meta-analysis of neurobehavioral outcomes in very preterm and/or very low birth weight children. Pediatrics 2009;124(2):717–28. https://doi.org/10.1542/peds.2008-2816.

[3] Hack M, Flannery DJ, Schluchter M, et al. Outcomes in young adulthood for very-low-birth-weight infants. N Engl J Med 2002;346(3):149–57. https://doi.org/10.1056/NEJMoa010856.

[4] de Jong F, Monuteaux MC, van Elburg RM, et al. Systematic review and meta-analysis of preterm birth and later systolic blood pressure. Hypertension 2012;59(2):226–34. https://doi.org/10.1161/HYPERTENSIONAHA.111.181784.

[5] Hornman J, De Winter AF, Kerstjens JM, et al. Emotional and behavioral problems of preterm and full-term children at school entry. Pediatrics 2016;137(5). https://doi.org/10.1542/peds.2015-2255.

[6] Kerr-Wilson CO, Mackay DF, Smith GCS, et al. Meta-analysis of the association between preterm delivery and intelligence. J Publ Health 2012;34(2):209–16. https://doi.org/10.1093/pubmed/fdr024.

[7] Kerstjens JM, De Winter AF, Bocca-Tjeertes IF, et al. Risk of developmental delay increases exponentially as gestational age of preterm infants decreases: a cohort study at age 4 years. Dev Med Child Neurol 2012;54(12):1096–101. https://doi.org/10.1111/j.1469-8749.2012.04423.x.

[8] Luu TM, Katz SL, Leeson P, et al. Preterm birth: risk factor for early-onset chronic diseases. CMAJ (Can Med Assoc J) 2016;188(10):736–46. https://doi.org/10.1503/cmaj.150450.

[9] Theunissen NC, den-Ouden AL, Meulman JJ, et al. Health status development in a cohort of preterm children. J Pediatr 2000;137(4):534–9. https://doi.org/10.1067/mpd.2000.108446.

[10] Walani SR. Global burden of preterm birth. Int J Gynecol Obstet 2020;150(1):31–3. https://doi.org/10.1002/ijgo.13195.

[11] March of Dimes PMNCH. In: Howson CP, Kinney MV, Lawn JE, editors. Save the children, WHO. Born too soon: the global action report on preterm birth. Geneva: World Health Organization; 2012 (n.d.).

[12] Tataranno ML. Early biomarkers of brain development in preterm infants. Doctoral dissertation. Utrecht University; 2018.

[13] Rees S, Inder T. Fetal and neonatal origins of altered brain development. Early Hum Dev 2005;81(9):753–61. https://doi.org/10.1016/j.earlhumdev.2005.07.004.

[14] Felderhoff-Mueser U, Bittigau P, Sifringer M, et al. Oxygen causes cell death in the developing brain. Neurobiol Dis 2004;17(2):273–82. https://doi.org/10.1016/j.nbd.2004.07.019.

[15] Noori S, Seri I. Hemodynamic antecedents of peri/intraventricular hemorrhage in very preterm neonates. Sem Fetal Neonatal Med 2015, August;20(No. 4):232–7. https://doi.org/10.1016/j.siny.2015.02.004. WB Saunders.

[16] Bellù R, de Waal K, Zanini R. Opioids for neonates receiving mechanical ventilation: a systematic review and meta-analysis. Arch Dis Child Fetal Neonatal Ed 2010;95(4):F241–51. https://doi.org/10.1136/adc.2008.150318.

[17] Smith GC, Gutovich J, Smyser C, et al. Neonatal intensive care unit stress is associated with brain development in preterm infants. Ann Neurol 2011;70(4):541–9. https://doi.org/10.1002/ana.22545.

[18] Boardman JP, Counsell SJ. Invited review: factors associated with atypical brain development in preterm infants: insights from magnetic resonance imaging. Neuropathol Appl Neurobiol 2020;46(5):413–21. https://doi.org/10.1111/nan.12589.

[19] Uchitel J, Vanhatalo S, Austin T. Early development of sleep and brain functional connectivity in term-born and preterm infants. Pediatr Res 2021:1–16. https://doi.org/10.1038/s41390-021-01497-4.

[20] Bennet L, Walker DW, Horne RS. Waking up too early–the consequences of preterm birth on sleep development. J Physiol 2018;596(23):5687–708. https://doi.org/10.1113/JP274950.

[21] Weisman O, Magori-Cohen R, Louzoun Y, et al. Sleep-wake transitions in premature neonates predict early development. Pediatrics 2011;128(4):706–14. https://doi.org/10.1542/peds.2011-0047.

[22] Scher MS. Topical review: understanding sleep ontogeny to assess brain dysfunction in neonates and infants. J Child Neurol 1998;13(10):467–74. https://doi.org/10.1177/088307389801301001.

[23] Scher MS, Steppe DA, Banks DL. Prediction of lower developmental performances of healthy neonates by neonatal EEG-sleep measures. Pediatr Neurol 1996;14(2):137–44. https://doi.org/10.1016/0887-8994(96)00013-6.

[24] van den Hoogen A, Teunis CJ, Shellhaas RA, et al. How to improve sleep in a neonatal intensive care unit: a systematic review. Early Hum Dev 2017;113:78–86. https://doi.org/10.1016/j.earlhumdev.2017.07.002.

[25] Werth J, Atallah L, Andriessen P, et al. Unobtrusive sleep state measurements in preterm infants–A review. Sleep Med Rev 2017;32:109–22. https://doi.org/10.1016/j.smrv.2016.03.005.

[26] Dereymaeker A, Pillay K, Vervisch J, et al. Review of sleep-EEG in preterm and term neonates. Early Hum Dev 2017;113:87–103. https://doi.org/10.1016/j.earlhumdev.2017.07.003.

[27] Curzi-Dascalova L, Mirmiran M. Manual of methods for recording and analyzing sleep-wakefulness states in preterm and full-term infant. Paris: Inserm; 1996. p. 1–160.

[28] Anders TF, Emde RN, Parmelee AH. A manual of standardized terminology, techniques and criteria for scoring of states of sleep and wakefulness in newborn infants. UCLA Brain Information Service/BRI Publications Office, NINDS Neurological Information Network; 1971.

[29] Bik A, Sam C, de Groot E. A scoping review of behavioral sleep stage classification methds for preterm infants. Sleep Med 2022;90:74–82. https://doi.org/10.1016/j.sleep.2022.01.006.

[30] Brazelton TB, Nugent JK. Neonatal behavioral assessment scale (No. 137). Cambridge University Press; 1995.

[31] Als H. Manual for the naturalistic observation of the newborn (preterm and full-term). Boston, MA: The Children's Hospital; 1981.

[32] Als H, Lester BM, Tronick EZ, et al. Manual for the assessment of the preterm infants' behavior (APIB). In: Fitzgerald HE, Lester BM, Yogman MW, editors. Theory and research in behavioral pediatrics, vol. 1. New York: Plenum Press; 1982. p. 65–131.

[33] Lester BM, Tronick EZ. The neonatal intensive care unit network neurobehavioral scale procedures. Pediatrics 2004;113(Supplement 2):641–67.

[34] Prechtl HF. The behavioural states of the newborn infant (a review). Brain Res 1974;76(2):185–212. https://doi.org/10.1016/0006-8993(74)90454-5.

[35] Stefanski M, Schulze K, Bateman D, et al. A scoring system for states of sleep and wakefulness in term and preterm infants. Pediatr Res 1984;18(1):58–62.

[36] Thoman EB. Sleeping and waking states in infants: a functional perspective. Neurosci Biobehav Rev 1990;14(1):93–107. https://doi.org/10.1016/S0149-7634(05)80165-4.

[37] Mirmiran M, Maas YG, Ariagno RL. Development of fetal and neonatal sleep and circadian rhythms. Sleep Med Rev 2003;7(4):321–34. https://doi.org/10.1053/smrv.2002.0243.

[38] Watanabe K. The neonatal electroencephalogram and sleep-cycle patterns. In: Eyre JA, editor. The neurophysiological examination of the newborn infant. New York: Mac Keith Press; 1992. p. 11–47.

[39] Liaw JJ, Yang L, Lo C, et al. Caregiving and positioning effects on preterm infant states over 24 hours in a neonatal unit in Taiwan. Res Nurs Health 2012;35(2):132–45. https://doi.org/10.1002/nur.21458.

[40] Knoop MS, de Groot ER, Dudink J. Current ideas about the roles of rapid eye movement and non–rapid eye movement sleep in brain development. Acta Paediatr 2021;110(1):36–44. https://doi.org/10.1111/apa.15485.

[41] de Groot ER, Knoop MS, van den Hoogen A, et al. The value of cardiorespiratory parameters for sleep state classification in preterm infants: a systematic review. Sleep Med Rev 2021:101462. https://doi.org/10.1016/j.smrv.2021.101462.

[42] Als H. Regarding the premature infant. Appendix B: behavioral definitions. In: Goldson E, editor. Nurturing the premature infant: developmental interventions in the neonatal intensive care nursery. Oxford: Oxford University Press; 1999. p. 77–85.

[43] Holditch-Davis D, Scher M, Schwartz T, et al. Sleeping and waking state development in preterm infants. Early Hum Dev 2004;80(1):43–64. https://doi.org/10.1016/j.earlhumdev.2004.05.006.

[44] Korner AF. State as variable, as obstacle, and as mediator of stimulation in infant research. Merrill-Palmer Q Behav Dev 1972;18(2):77–94.

[45] Wolff PH. The causes, controls, and organization of behavior in the neonate. Psychol Issues 1966;5(1):1–105.

[46] Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971;76(5):378. https://doi.org/10.1037/h0031619.

[47] Field A. Discovering statistics using IBM SPSS statistics. Sage; 2013.

[48] Prechtl HFR, O'Brien MJ. Behavioural states of the full-term newborn: the emergence of a concept. In: Stratton P, editor. Psychobiology of the human newborn. New York: John Wiley and Sons Ltd; 1982. p. 53–73.

[49] Hayes MJ, Plante LS, Fielding BA, et al. Functional analysis of spontaneous movements in preterm infants. Dev Psychobiol: J Int Soc Develop Psychobiol 1994;27(5):271–87. https://doi.org/10.1002/dev.420270503.

[50] Thoman EB. Early development of sleeping behaviors in infants. In: Ellis NR, editor. Aberrant development in infancy. Routledge; 1975. p. 123–38.

[51] Curzi-Dascalova L, Peirano P, Morel-Kahn F. Development of sleep states in normal premature and full-term newborns. Dev Psychobiol: J Int Soc Develop Psychobiol 1988;21(5):431–44. https://doi.org/10.1002/dev.420210503.

[52] Bourel-Ponchel E, Hasaerts D, Challamel MJ, et al. Behavioral-state development and sleep-state differentiation during early ontogenesis. Neurophysiol Clin 2021;51(1):89–98. https://doi.org/10.1016/j.neucli.2020.10.003.