Published in final edited form as:

Nurs Res. 2017 ; 66(4): 323–329. doi:10.1097/NNR.0000000000000224.

Maintaining Interrater Agreement of Core Assessment Instruments in a Multisite Randomized Controlled Clinical Trial: The Randomized Evaluation of Sedation Titration for Respiratory Failure (*RESTORE*) Trial

Ruth Lebet, MSN, CCNS-P,

Lecturer, School of Nursing, University of Pennsylvania, Philadelphia, PA

Jennifer Hayakawa, DNP, PCNS-BC, CNRN, CCRN,

Clinical Nurse Specialist, PICU, CHOC Children's Hospital, Orange, CA and Clinical Faculty, Western University of Health Sciences, Pomona, CA

Tracy B. Chamblee, PhD, APRN, PCNS-BC,

Clinical Nurse Specialist, PICU, Children's Medical Center Dallas, Dallas, TX

Joana A. Tala, MD,

Research Coordinator, Pediatric Intensive Care Unit, Yale New Haven Hospital/Yale University, New Haven, CT

Nakul Singh, ScM,

Biostatistician, Department of Cardiology, Boston Children's Hospital, Boston, MA

David Wypij, PhD, and

Senior Biostatistician, Department of Cardiology, Boston Children's Hospital and Associate Professor, Department of Pediatrics, Harvard Medical School and Senior Lecturer, Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA

Martha A. Q. Curley, RN, PhD, FAAN

Ellen and Robert Kapito Professor in Nursing Science, School of Nursing, University of Pennsylvania, Philadelphia, PA, and Nurse Scientist, Boston Children's Hospital, Boston, MA

Abstract

Corresponding Author: Ruth Lebet, University of Pennsylvania School of Nursing, Claire M. Fagin Hall, 418 Curie Boulevard, Philadelphia, PA 19104-4217 lebet@nursing.upenn.edu.

Ruth Lebet, MSN, CCNS-P, is Lecturer, School of Nursing, University of Pennsylvania, Philadelphia, PA.

Jennifer Hayakawa, DNP, PCNS-BC, CNRN, CCRN, is Clinical Nurse Specialist, PICU, CHOC Children's Hospital, and Clinical Faculty, Western University of Health Sciences.

Tracy B. Chamblee, PhD, APRN, PCNS-BC, is Clinical Nurse Specialist, PICU, Children's Medical Center Dallas.

Joana A. Tala, MD, is Research Coordinator, Pediatric Intensive Care Unit, Yale New Haven Hospital/Yale University.

Nakul Singh, ScM, is Biostatistician, Department of Cardiology, Boston Children's Hospital.

David Wypij, PhD, is Senior Biostatistician, Department of Cardiology, Boston Children's Hospital, and Associate Professor, Department of Pediatrics, Harvard Medical School, and Senior Lecturer, Department of Biostatistics, Harvard T. H. Chan School of Public Health.

Martha A. Q. Curley, RN, PhD, FAAN, is Ellen and Robert Kapito Professor in Nursing Science, School of Nursing, University of Pennsylvania, and Nurse Scientist, Boston Children's Hospital.

This paper is a supplementary analysis of data collected during the Randomized Evaluation of Sedation Titration for Respiratory Failure (*RESTORE*) clinical trial. Trial Registration: clinicaltrials.gov, Identifier: NCT00814099.

The authors have no conflicts of interest to report.

Background—*RESTORE* (Randomized Evaluation of Sedation Titration for Respiratory failure) was a cluster randomized clinical trial evaluating a sedation strategy in children 2 weeks to < 18 years of age with acute respiratory failure supported on mechanical ventilation. A total of 31 U.S. pediatric intensive care units (PICUs) participated in the trial. Staff nurse rater agreement on measures used to assess a critical component of treatment fidelity was essential throughout the four-year data collection period.

Objective—The purpose of the paper is to describe the method of establishing and maintaining interrater agreement (IRA) of two core clinical assessment instruments over the course of the clinical trial.

Methods—IRA cycles were carried out at all control and intervention sites and included a minimum of five measurements of the State Behavioral Scale (SBS) and Withdrawal Assessment Tool-Version 1 (WAT-1). Glasgow Coma Scale scores were also obtained. PICUs demonstrating < 80% agreement repeated their IRA cycle. Fleiss's kappa coefficient was used to assess interrater agreement.

Results—Repeated IRA cycles were required for 8% of 226 SBS cycles and 2% of 222 WAT-1 cycles. Fleiss's kappa coefficients from more than 1,350 paired assessments were 0.86 for SBS and 0.92 for WAT-1, demonstrating strong agreement and similar to 0.91 for the Glasgow Coma Scale. There was no difference in Fleiss's kappa for any of the instruments based on unit size or timing of assessment (earlier or later in the study). For SBS scores, Fleiss's kappa was significantly different in larger and smaller PICUs (0.82 vs. 0.92, $p < .003$); however, Fleiss's kappa for both groups indicated excellent agreement.

Conclusion—Monitoring measurement reliability is an essential step in ensuring treatment fidelity and, thus, the validity of study results. Standardization on the use of these core assessment instruments among participating sites was achieved and maintained throughout the trial.

Keywords

interrater agreement; pediatric intensive care; State Behavioral Scale; sedation assessment; treatment fidelity; Withdrawal Assessment Tool-Version 1; withdrawal assessment

With an increasing focus on evidence-based practice, the impact of implementing a well-designed randomized controlled trial (RCT) is more important than ever. Although RCTs remain the gold standard for testing interventions, inadequate methodological approaches to RCT planning and implementation can lead to biased results, as well as inaccurate interpretation of study findings (Altman et al., 2001; Moher, Schulz, & Altman, 2001). When protocol steps are based on assessment findings and members of the care team do not use assessment tools correctly, protocol implementation is compromised. This can result in a large standard error when effect size is calculated, indicating less precision (Page, 2014).

Treatment fidelity is monitored to ensure that a planned intervention is implemented as designed in an RCT. Treatment fidelity is defined as methodological strategies used to monitor and enhance the reliability and validity of a clinical intervention (Bellg et al., 2004). Monitoring treatment fidelity is a strategy to ensure the validity of study findings. When a study protocol includes or is based on clinical assessment tools, it is important to ensure

measurement reliability as an important component of treatment fidelity. The Treatment Fidelity Workgroup of the National Institutes of Health Behavioral Change Consortium stressed the importance of standardized training, measurement of skill acquisition, and minimizing decay or change in skills over time (Borrelli, 2011). Consistent with those recommendations, we describe strategies used to ensure measurement reliability—indicated by rater agreement—of two assessment tools as one aspect of treatment fidelity in a clinical trial implemented in a pediatric critical care setting.

Background

A ventilated patient's sedation needs are highly individualized and dependent on multiple factors, such as the nature and course of illness, drug interactions, and concomitant therapies. To optimize patient comfort in the pediatric intensive care unit (PICU) setting, sedatives are commonly used in conjunction with analgesics; these drugs are administered via intermittent or continuous infusion. Undersedation is associated with unresolved pain and ineffective ventilation, as well as self-removal of critical invasive lines and tubes (Grant et al., 2012). Oversedation is equally detrimental and is associated with potential iatrogenic effects such as tolerance, withdrawal syndrome, pressure ulcers, and ventilator associated pneumonia (Vet et al., 2013).

The *RESTORE* (Randomized Evaluation of Sedation Titration fOr Respiratory failurE) study was a cluster randomized clinical trial designed to test a novel approach to sedation management in pediatric patients 2 weeks to < 18 years of age with acute respiratory failure supported on mechanical ventilation (Curley et al., 2015). Thirty-one PICUs in the US participated in the trial. The trial intervention consisted of daily assessment of the mechanically ventilated patient's illness trajectory, establishment of an individualized sedation goal, and implementation of a nurse-directed comfort algorithm that guided the sedation/analgesic management. The intervention required daily prescription of a sedation goal using a standardized pediatric sedation assessment instrument. In addition, iatrogenic withdrawal syndrome (IWS) was prospectively monitored as an adverse event in all at-risk subjects. The *RESTORE* trial was approved by the Institutional Review Board (IRB) of each participating center. Parental permission was obtained for each enrolled child. When feasible, assent was also obtained from children 8 years and older.

The trial identified that use of a nurse-managed sedation protocol did not reduce duration of mechanical ventilation among children receiving mechanical ventilation for acute respiratory failure when compared with usual care. Importantly, analyses of secondary outcomes identified that compared with patients receiving usual care, patients in the intervention group were safely managed in a more awake and calm state while intubated, had fewer days of opioid exposure, and received drugs from fewer sedative classes (Curley et al., 2015).

Given that data were collected over a four-year period of time, the potential for secular changes in sedation and IWS assessments was significant. The *RESTORE* clinical trial targeted sedation using the State Behavioral Scale (SBS). This pediatric-specific, validated sedation assessment instrument describes the sedation-agitation continuum experienced by intubated patients supported on mechanical ventilation (Curley, Harris, Fraser, Johnson, &

Arnold, 2006). The sedation-agitation continuum ranges over six levels from $-3 = \textit{unresponsive}$ to $+2 = \textit{agitated}$. Use of the instrument provides data that are simple to calculate and document, accurately describes the degree of sedation, and was used in the *RESTORE* intervention to guide nurse-managed sedation/analgesia therapy.

Patients exposed to five or more days of opioid and/or sedation therapy are at risk for the development of IWS. Abrupt discontinuation or rapid weaning of opioids or benzodiazepines is associated with IWS that includes central nervous system (CNS), gastrointestinal, and motor symptoms that may negatively impact the patient's recovery and time to discharge. The Withdrawal Assessment Tool–Version 1 (WAT-1) is a validated 11-item, 12-point scale designed to quantify the presence and severity of withdrawal symptoms in children at risk for development of IWS (Franck, Harris, Soetenga, Amling, & Curley, 2008; Franck, Scoppettuolo, Wypij, & Curley, 2011). The WAT-1 was used in the *RESTORE* trial to facilitate early identification of IWS, as well as response to treatment.

Each of these scores can be grouped into clinically relevant categories. In our process, scores for each assessment were grouped into clinically relevant categories when data was analyzed. SBS scores were grouped into four categories: $-3 = \textit{unresponsive}$; $-2 = \textit{responsive to noxious stimuli}$; -1 or $0 = \textit{awake}$; and 1 or $2 = \textit{agitated}$. In a similar fashion, WAT-1 scores were grouped into two categories: $(0-2) = \textit{no withdrawal present}$ and $3 = \textit{withdrawal present}$. GCS scores were grouped into three categories according to different levels of consciousness as $(3-8) = \textit{coma}$, $(9-12) = \textit{lethargy}$, and $(13-15) = \textit{awake}$.

Methods

SBS and WAT-1 Training

Prior to implementation of the *RESTORE* clinical trial, nurses and physicians at each of the participating sites (assigned randomly to intervention or control status) agreed to implement the SBS and WAT-1 assessment tools as a unit-based standard of care. All PICUs received standardized training on the use of the SBS and WAT-1 instruments. The training included a voice-over PowerPoint presentation with patient videos, allowing learners to practice using the assessment tools on standardized patients. Nursing and physician staff at each site then completed a 15-item posttest. Participants scoring $< 100\%$ on the posttest received individualized review of the content with the nurse co-investigator at the site, followed by retesting on the content until mastery was documented. As new staff was brought on board, they received the same training. Participating sites were also provided with a *RESTORE* training manual and access to the *RESTORE* website for additional resources.

Interrater Agreement Cycles

Upon completion of initial training, initial interrater agreement (IRA) for each site was established and then routinely assessed throughout the four-year clinical trial. The IRA cycle process is shown in Figure 1.

The assessment cycles were initially performed approximately three times/year at all participating sites. As the instruments became integrated into practice and the IRA stabilized, the IRA cycle interval was widened to twice a year. A standard operating

procedure for IRA assessment was developed and provided to each site with each request for completion of an IRA cycle.

Two strategies were used to ensure representativeness of the unit staff during the IRA cycle. Nurses who rotated to night shift completed at least two of the five required paired assessments. In addition, each request for an IRA cycle was sent to the site nurse co-investigator on a randomly selected day with a request for completion within 48 hours.

To perform an IRA cycle, the nurse co-investigator first identified patients appropriate to be assessed. PICU census and patient acuity is variable throughout the year, so this standard method of patient selection, based on prioritized criteria, was used by all sites. Patients enrolled in the *RESTORE* trial were preferred, so enrolled study patients were chosen first for assessment. If there were no currently enrolled study patients, patients on continuous infusions of sedation/pain medications who were receiving these assessments as a unit-based standard of care were assessed. If no patients met either of these criteria, patients on intermittent doses of sedation/pain medications were assessed. This process was approved by all sites' IRBs.

Next, the nurse co-investigator invited nurses who were trained in the use of the *RESTORE* protocol and who could be freed up from their patient assignment to participate in the IRA cycle. Generally one of the nurses in the pair was the nurse caring for the patient being assessed. Nurse pairs completed concurrent independent assessments on sedated intubated ventilated patients in their PICU. It was preferred that 10 different nurses make up the five pairs but, if necessary, the nurse co-investigator would pair with five different nurses to complete five pairs of assessments. The nurse raters recorded their scores on separate rater scorecards and were blind to each other's assessments at all times. A minimum of five separate pairs of assessments of the SBS and WAT-1 instruments was required during each IRA cycle, although sites were encouraged to complete more. In a single IRA cycle, five different pairs of nurse raters were required, but it was possible that a nurse rater could participate in more than one IRA cycle.

During IRA cycles the Glasgow Coma Scale (GCS) was scored at the same time the SBS was scored. Since its development in 1974 (Teasdale & Jennett, 1974), the GCS has been considered the gold standard for assessing a patient's level of consciousness. Because of its simplicity and demonstrated effectiveness with assessing neurologic findings, assessment of GCS scoring was completed concurrently with SBS scoring as a means to evaluate the construct validity of the SBS by comparison to an alternative instrument measuring level of consciousness. State profiles generated on the basis of SBS dimensions (respiratory drive/response to ventilation, coughing, response to stimulation, attentiveness to care provider, tolerance to care, consolability, and movement after consoled) were validated by comparing group mean scores to a numeric rating scale for sedation/agitation (Curley, Harris, Fraser, Johnson, & Arnold, 2006).

Once obtained, raw IRA scores were sent to the Data Coordinating Center for analysis. PICUs demonstrating < 80% agreement were required to remediate by repeating their IRA

cycle. Any unit failing a remediation IRA cycle was required to implement a unitwide quality improvement plan and then repeat the IRA cycle.

An additional strategy was used to assess for local differences in assessment tool use between PICUs, which can be an issue in multicenter clinical trials. A quality monitoring plan—developed prior to study initiation—specified periodic visits by quality monitors trained in the use of the IRA cycle assessment tools. The quality monitors performed SBS and WAT-1 assessments with three staff nurses at each site during each site visit. Any issues in assessment tool use were immediately addressed by the quality monitor, including re-education.

Although not part of the IRA cycle, documentation of the assessment tool scores was also tracked as another aspect of treatment fidelity, adherence to treatment protocols. This information can be found in the parent paper as Supplemental eTable 5: Assessment Adherence per Eligible Patient-Days According to Group (Curley et al., 2015).

Data Analysis

IRA was evaluated via simple concordance rates and using Fleiss's kappa coefficient (Fleiss, 1971), a measure of nominal scale agreement between pairs of raters when there are multiple raters and agreement due to chance is factored out. To evaluate the IRA, scores for each assessment were grouped into the previously described, clinically relevant categories when data was analyzed: SBS scores were grouped as *unresponsive* (−3), *responsive to noxious stimuli* (−2), *awake* (−1, 0), and *agitated* (1, 2); WAT-1 scores were grouped as *no withdrawal present* (0–2), and *withdrawal present* (3); and GCS scores were grouped into three categories according to different levels of consciousness as *coma* (3–8), *lethargy* (9–12), and *awake* (13–15).

To be conservative, data from repeated IRA cycles after a site had failed to demonstrate 80% agreement were not included in findings reported here. Construct validity of the SBS in comparison to the GCS was evaluated using Spearman's rank correlation coefficient, restricting to the subset of patients scored as nonagitated (SBS = 0). Fleiss's kappa coefficients were also compared across subgroups, including the randomized site-specific treatment group assignment (intervention versus control), PICU size (large, with > median number of RN staff, versus small), and assessment timing (early [within the first four IRA cycles] vs. late).

Results

PICU Characteristics

Organizational assessment data for the 31 PICUs in the *RESTORE* trial are presented in Table 1. The majority of PICUs were affiliated with an academic medical center and about half had Magnet Recognition (American Nurses Credentialing Center, Silver Spring, MD). The median number of RN staff across PICUs was 69. Of the 31 PICUs, 17 were randomized to the intervention group and 14 to the control group. Sites were categorized as small (< median number of RN staff) or large (> median number of RN staff). More detailed

information is available in Curley et al. (2015, eTable 1: Prerandomization Characteristics of the 31 Participating PICUs According to Size).

Among the 22 PICUs participating when the *RESTORE* trial began, the number of IRA cycles ranged from 7–9 ($M = 8.7$). To increase enrollment in the trial, nine additional PICUs began participating approximately one year later; the number of IRA cycles ranged from 3–4 ($M = 3.9$) in this group. Among 226 SBS cycles across the 31 PICUs, 18 cycles (8%) had to be repeated across 14 PICUs. Two sites had to repeat an SBS cycle three times. Among 222 WAT-1 cycles across the 31 PICUs, four cycles (2%) had to be repeated across three PICUs. One site had to repeat a WAT-1 cycle twice. In no case did a site fail a repeated IRA cycle, so no sites were required to implement a unitwide quality improvement plan.

Interrater Agreement

Interrater agreement is summarized in Table 2. After excluding observations from repeated cycles, the 1,446 pairs of SBS observations had a concordance rate of 92% and Fleiss's kappa coefficient of 0.86, indicating strong agreement in scoring among the nurse raters. Similarly, the 1,357 pairs of WAT-1 observations had a concordance rate of 97% and kappa coefficient of 0.92, again indicating strong agreement in scoring among the nurse raters. These measures of agreement matched well with those of the GCS, as the 1,408 pairs of GCS observations had a concordance rate of 95% and kappa coefficient of 0.91. Restricting to 2,521 paired SBS and GCS observations where the nurse rater scored the patient as nonagitated (i.e., SBS = 0), there were strong associations between the SBS and GCS scores (for original scores, $r_s = .74$, $p < .001$; for grouped scores, $r_s = .63$, $p < .001$).

To determine if high IRA existed among subgroups, we compared kappa coefficients of intervention and control sites, large and small PICUs, and early versus late assessment timing. Findings are summarized in Table 3. When comparing intervention and control sites, and early versus late assessment timing, the agreement remained very high, with no statistically significant differences comparing any of the subgroups for SBS, WAT-1, or GCS. There was a statistically significant difference in the Fleiss's kappa for the SBS when large and small PICUs were compared, although the kappa for both groups indicated excellent agreement.

Discussion

This report provides an in-depth examination of one aspect of treatment fidelity, rater agreement on scoring of assessment tools used to drive a nurse-managed sedation protocol. Assessment drives intervention, so ensuring accuracy and consistency in assessment supports study findings and facilitates appropriate care.

In this multicenter trial, monitoring the agreement among pairs of nurse's assessments using the SBS, WAT-1, and GCS instruments was implemented as one of a series of treatment fidelity strategies as described by Bellg et al. (2004). Specifically, this strategy was used to monitor and improve enactment of treatment skills to allow confident comparisons of the level of sedation and IWS across treatment groups. Additionally, for the intervention group, the SBS and WAT-1 scores were used to drive the treatment protocol for each patient. The

primary goal of treatment fidelity is to reduce variation in the delivery of the study intervention. This cannot be achieved if key clinical assessments used to implement the protocol are not accurate and consistent. Despite many differences among participating sites, high agreement on key assessments required for implementation of the study intervention was established and maintained over the course of the trial, supporting treatment fidelity, and consequently supporting study findings. eTable 6: Protocol Compliance in the PICU in Supplement 1 of the parent paper provides detailed information about treatment fidelity achieved in the *RESTORE* trial (Curley et al., 2015). In brief, adherence to core elements of the sedation protocol ranged from 71% to 100% of eligible study days and from 86% to 98% of patients. Intervention sites demonstrated high fidelity (>80% compliance [Borrelli, 2011]) on eight of the ten core protocol elements and moderate fidelity on two elements.

In addition, ease of use and expanded validity of both instruments was demonstrated in a dynamic and complex patient care setting. The SBS score, originally validated in children 6 weeks to 6 years of age (Curley, Harris, Fraser, Johnson, & Arnold, 2006), demonstrated good agreement and construct validity in patients 2 weeks to < 18 years of age. This is indicated by the high concordance rate and Fleiss's kappa coefficient among nurse raters, as well as strong association between the SBS and GCS scores.

Limitations

The IRA cycle process had some limitations. The cluster randomized design may have allowed local differences in units to influence scoring. However, in general, comparison across subgroups showed high levels of agreement. This was supported by the findings of quality monitors who performed SBS and WAT-1 assessments with three staff nurses at each site during periodic site visits.

Scoring differences were also seen (refer to Tables 2 and 3). Comparison of small versus large units showed a statistically significant difference. Large units, with more nursing staff, generally involved more study staff in the educational process, which may have resulted in some increased variability in assessment tool use for the SBS. One of 1,446 SBS observations had a difference of two categories, with one nurse scoring the patient as *agitated* and the second as *unresponsive*. For the WAT-1, six of 1,357 observations showed a difference of two categories (no withdrawal vs. withdrawal). Although the IRA cycle procedure stated that scores were to be obtained simultaneously, it is possible that the scores were obtained a few minutes apart. It is also possible that nurses participating in an IRA assessment who cared for the patient may have been influenced by prior knowledge of the patient.

Implications for Future Research

The success of the strategies used in this trial has several implications for future intervention trials. First, creation of a standardized education plan prior to study implementation provided the foundation and set the expectation for the study sites regarding these key assessments. To do this, we used voice-over PowerPoint and videos providing standardized patients for all study sites to assess and score. Training was reinforced as needed based on posttest scores. Next, an IRA cycle plan was clearly defined and documented in a standard

operating procedure. The plan and process were communicated to study sites prior to study implementation. Further, a minimal level of agreement (80%) for all sites was established and communicated as an expectation. Sites that failed to achieve this threshold were charged with developing and implementing an improvement plan, after which performance was re-evaluated. Finally, to provide ongoing feedback and minimize practice drift in tool use, IRA cycle results were routinely discussed with participating sites.

Nurse-investigators planning a clinical trial at one or multiple sites—which requires repeated clinical assessments using scoring tools—can use this discussion as a template for designing a plan to reduce unintended variability in tool use. Key elements of the plan include developing a standardized education plan, and detailing the plan for ensuring interrater agreement as part of protocol development. A detailed, written, standard operating procedure for the process which includes easily understood forms, establishment of a target level of agreement, and a plan for remediation when targets are not met are crucial elements. The plan utilized should be included in research reports.

Conclusions

In summary, we report consistent strong interrater agreement scores on the SBS and WAT-1 assessment tools from more than 1,350 paired assessments over a four-year-period across 31 clinical sites. The SBS and WAT-1 were key assessments required for the nurse-driven intervention tested in the *RESTORE* trial; the potential impact of variation in these key assessments on the trial intervention was significant. Use of a standardized training and monitoring plan was instrumental in our efforts to achieve high agreement on the core assessments and enhance the overall quality of the study outcomes.

Acknowledgments

This article was supported, in part, by the National Institutes of Health, the National Heart, Lung, and Blood Institute, and the National Institute of Nursing Research: U01 HL086622 and U01 HL086649; PI: Curley & Wypij.

The authors gratefully acknowledge the statistical support of Lisa A. Asaro, MS, Biostatistician II, and IRA cycle coordination support of Donna M. Duva, Project Manager, Data Coordinating Center, Department of Cardiology, Children's Hospital, Boston. In addition, they acknowledge the support of the *RESTORE* Investigators and site teams.

References

- Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, ... Lang T. The revised CONSORT statement for reporting randomized trials: Explanation and elaboration. *Annals of Internal Medicine*. 2001; 134:663–694. DOI: 10.7326/0003-4819-134-8-200104170-00012 [PubMed: 11304107]
- Bellg AJ, Borrelli B, Resnick B, Hecht J, Minicucci DS, Ory M, ... Czykowski S. Enhancing treatment fidelity in health behavior change studies: Best practices and recommendations from the NIH Behavior Change Consortium. *Health Psychology*. 2004; 23:443–451. DOI: 10.1037/0278-6133.23.5.443 [PubMed: 15367063]
- Borelli B. The assessment, monitoring, and enhancement of treatment fidelity in public health clinical trials. *Journal of Public Health Dentistry*. 2011; 71:S52–S63. DOI: 10.1111/j.1752-7325.2011.00233.x
- Curley MAQ, Harris SK, Fraser KA, Johnson RA, Arnold JH. State Behavioral Scale: A sedation assessment instrument for infants and young children supported on mechanical ventilation. *Pediatric*

- Critical Care Medicine. 2006; 7:107–114. DOI: 10.1097/01.PCC.0000200955.40962.38 [PubMed: 16446601]
- Curley MAQ, Wypij D, Watson R, Grant MJC, Asaro LA, Cheifetz IM, ... Matthay MA. Protocolized sedation vs. usual care in pediatric patients mechanically ventilated for acute respiratory failure: A randomized clinical trial. JAMA. 2015; 313:379–389. DOI: 10.1001/jama.2014.18399 [PubMed: 25602358]
- Fleiss JL. Measuring nominal scale agreement among many raters. Psychological Bulletin. 1971; 76:378–382. DOI: 10.1037/h0031619
- Franck LS, Harris SK, Soetenga DJ, Amling JK, Curley MAQ. The Withdrawal Assessment Tool-Version 1 (WAT-1): An assessment instrument for monitoring opioid and benzodiazepine withdrawal symptoms in pediatric patients. Pediatric Critical Care Medicine. 2008; 9:573–580. DOI: 10.1097/PCC.0b013e31818c8328 [PubMed: 18838937]
- Franck LS, Scoppettuolo LA, Wypij D, Curley MAQ. Validity and generalizability of the Withdrawal Assessment Tool-1 (WAT-1) for monitoring iatrogenic withdrawal syndrome in pediatric patients. Pain. 2012; 153:142–148. DOI: 10.1016/j.pain.2011.10.003 [PubMed: 22093817]
- Grant MJC, Scoppettuolo LA, Wypij D, Curley MAQ. Prospective evaluation of sedation-related adverse events in pediatric patients ventilated for acute respiratory failure. Critical Care Medicine. 2012; 40:1317–1323. DOI: 10.1097/CCM.0b013e31823c8ae3 [PubMed: 22425823]
- Moher D, Schulz KF, Altman DG. The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet. 2001; 357:1191–1194. DOI: 10.1016/S0140-6736(00)04337-3 [PubMed: 11323066]
- Page P. Beyond statistical significance: Clinical interpretation of rehabilitation research literature. International Journal of Sports Physical Therapy. 2014; 9:726–736. [PubMed: 25328834]
- Teasdale G, Jennett B. Assessment of coma and impaired consciousness: A practical scale. Lancet. 1974; 304:81–84. DOI: 10.1016/S0140-6736(74)91639-0
- Vet NJ, Ista E, de Wildt SN, van Dijk M, Tibboel D, de Hoog M. Optimal sedation in pediatric intensive care patients: A systematic review. Intensive Care Medicine. 2013; 39:1524–1534. DOI: 10.1007/s00134-013-2971-3 [PubMed: 23778830]

RESTORE Interrater Agreement Cycle (IRAC) Process

FIGURE 1.
The interrater agreement cycle (IRAC) process used in the *RESTORE* trial.

TABLE 1Characteristics of PICUs Participating in the *RESTORE* Clinical Trial

Characteristic	<i>n</i>	(%)
Academic medical center (yes)	28	90
Magnet designation (yes)	15	48
	<i>Mdn</i>	(<i>Q1</i> , <i>Q3</i>)
PICU beds (number)	22	(16, 29)
Annual admissions (number)	1116	(792, 1827)
Daily census (average)	14.5	(11.0, 19.1)
RN staff (number)	69	(47, 97)
Full-time RNs (%)	77	(65, 86)
BSN-prepared (%) ^a	80	(74, 90)
PICU nursing experience (years) (average) ^b	6.2	(5.1, 8.3)
Overall nursing experience (years) (average) ^b	7.8	(6.0, 10.0)
CNS or unit-based educators (FTEs) (number)	1	(1, 2)

Note. *N* = 31. BSN = bachelor of science degree in nursing; CNS = clinical nurse specialist; FTE = full time equivalent; *Mdn* = median; PICU = pediatric intensive care unit; *Q1* = 25th percentile; *Q3* = 75th percentile; RN = registered nurse.

^aBSN-prepared unavailable for two sites.

^bNursing experience unavailable for nine sites.

TABLE 2
Interrater Agreement: State Behavioral Scale, Glasgow Coma Scale, and Withdrawal Assessment Tool–Version 1

Scale/score group	Score group				Concordance		Fleiss's κ	
	–3	–2	–1, 0	+1, +2	Est.	95% CI	Est.	95% CI
State Behavioral Scale					.92	[.91, .94]	.86	[.83, .90]
–3 Unresponsive	124							
–2 Responsive to noxious stimuli	15	261						
–1, 0 Awake	2	56	819					
+1, +2 Agitated	0	1	40	128				
Glasgow Coma Scale					.95	[.93, .96]	.91	[.88, .95]
3–8 Coma	507							
9–12 Lethargy	51	571						
13–15 Awake	7	18	254					
WAT–1					.97	[.96, .98]	.92	[.86, .97]
0–2 No withdrawal	967							
3–12 Yes withdrawal	45	345						

Note. Cell entries for score groups are counts. CI = confidence interval; est. = estimate; WAT–1 = Withdrawal Assessment Tool–Version 1.

TABLE 3Interrater Agreement (Fleiss' κ) by Subgroups

Subgroup	State Behavioral Scale					WAT-1					Glasgow Coma Scale				
	<i>n</i>	κ	95% CI	<i>p</i>	<i>n</i>	κ	95% CI	<i>p</i>	<i>n</i>	κ	95% CI	<i>p</i>	<i>n</i>	κ	95% CI
Treatment group				.76				.63				.26			
Control	599	.86	[.80, .91]		581	.93	[.85, 1.00]		574	.94	[.88, 1.00]		574	.94	[.88, 1.00]
Intervention	847	.87	[.82, .91]		776	.90	[.83, .97]		834	.90	[.85, .95]		834	.90	[.85, .95]
Unit size				.003				.92				.30			
Large	801	.82	[.77, .86]		753	.92	[.85, .99]		796	.90	[.84, .95]		796	.90	[.84, .95]
Small	645	.92	[.87, .97]		604	.91	[.83, .99]		612	.94	[.88, .99]		612	.94	[.88, .99]
Assessment timing				.60				1.00				.26			
Early	631	.85	[.80, .90]		564	.92	[.83, 1.00]		601	.89	[.83, .95]		601	.89	[.83, .95]
Late	815	.87	[.83, .92]		793	.92	[.85, .99]		807	.93	[.88, .98]		807	.93	[.88, .98]

Note. CI = confidence interval; WAT-1 = Withdrawal Assessment Tool-Version 1.