Evaluation of Interobserver Agreement of Apgar Scoring in Preterm Infants

WHAT’S KNOWN ON THIS SUBJECT: The Apgar score is a convenient method to rapidly assess the clinical status of the newborn infant. Recent literature suggests Apgar scores vary widely in preterm infants.

WHAT THIS STUDY ADDS: The Apgar signs for respiratory effort, grimace, and muscle tone demonstrated considerable disagreement in preterm infants ≤28 weeks’ gestation. Disagreement exists despite the level of respiratory intervention, continuous positive airway pressure, or intubation, and is likely independent of gestational age.

abstract

BACKGROUND AND OBJECTIVES: Apgar scoring is accepted by medical professionals both as a measure of the infant’s clinical status and the infant’s response to resuscitation. Recent studies, however, have suggested significant variability when used for scoring preterm infants. We hypothesized that agreement in Apgar scoring would improve with increasing gestational age and at low levels of respiratory support. We also hypothesized that grimace and muscle tone would demonstrate the least agreement.

METHODS: Neonatologists from the Perinatal Section of the American Academy of Pediatrics were presented with 4 film clip cases via a secure online survey: (1) full-term infant in room air; (2) 28 weeks’ gestation infant with continuous positive airway pressure; (3) 28 weeks’ gestation infant intubated; and (4) 24 weeks’ gestation infant intubated. Participants were shown 30-second clips at 1, 5, and 10 minutes of life and were asked to provide Apgar scores. κ coefficients were used to compare agreement for each component.

RESULTS: A total of 335 neonatologists participated in the survey. κ coefficients in the full-term infant for respiratory effort (0.94, 0.91), grimace (0.91, 0.90), and muscle tone (0.91, 0.89) demonstrated almost perfect agreement at 1 and 5 minutes. For preterm infants, respiratory effort (range: 0.07–0.40), muscle tone (range: 0.10–0.75), and grimace (range: 0.11–0.71) all demonstrated disagreement at 1, 5, and 10 minutes of life unless the infants were apneic and limp.

CONCLUSIONS: An improved delivery room score that decreases variability among medical care professionals is needed to accurately reflect the clinical status of preterm infants. Pediatrics 2012;130:e982–e987

AUTHORS: Monuj T. Bashambu, MD,a Halana Whitehead, BA,b Anna Maria Hibbs, MD, MSCE,a,b Richard J. Martin, MD,a,b and Monika Bhola, MDa,b

“Department of Pediatrics, Rainbow Babies and Children’s Hospital, Cleveland, Ohio; and Case Western Reserve University School of Medicine, Cleveland, Ohio

KEY WORDS agreement, Apgar score, preterm infants

ABBREVIATIONS CPAP—continuous positive airway pressure
ET PPV—endotracheal positive pressure ventilation
PPV—positive pressure ventilation

www.pediatrics.org/cgi/doi/10.1542/peds.2012-0368
doi:10.1542/peds.2012-0368
Accepted for publication May 24, 2012
Address correspondence to Monuj T. Bashambu, MD, Department of Pediatrics, Division of Neonatology, Rainbow Babies and Children’s Hospital, 11100 Euclid Ave, RBC Suite 3100, Cleveland, OH 44106. E-mail: monuj.bashambu@case.edu

PEDIATRICS (ISSN Numbers: Print, 0031-4005; Online, 1098-4275). Copyright © 2012 by the American Academy of Pediatrics

FINANCIAL DISCLOSURE: The authors have no financial relationships relevant to this article to disclose.

FUNDING: No external funding.
The first assessment an infant receives in the delivery room is supplemented by the assignment of Apgar scores. In 1952, Virginia Apgar introduced an infant scoring system at the 27th Annual Congress of Anesthetists as a method for comparison “of the results of obstetric practices, types of maternal pain relief and the effects of resuscitation.” Her efforts culminated in a new score, the “APGAR” score, evaluating color, heart rate, grimace, muscle tone, and respiratory effort. After ~60 years in practice, the Apgar score remains widely used by clinicians, nurses, and other neonatal caregivers. It serves both as a measure of the infant’s clinical status as well as a measure of the infant’s response to resuscitation. Various studies have attempted to link low Apgar scores to infants at increased risk for mortality and cerebral palsy.

Despite its international recognition and use in the delivery room, the score’s value in preterm infants has been challenged. Several small studies have shown variation in scoring of preterm infants, particularly when artificial ventilation is instituted in the first 10 minutes of life. O’Donnell et al previously described variation in the 5-minute Apgar score, both in term and preterm infants, when participants were shown 10-second video clips. This variation occurred regardless of illness severity and did not differ among groups of medical and nursing staff. In a separate study, O’Donnell et al also showed significant variation in color even when resuscitation was not needed. Finally, a review of 255 Polish neonatal centers found that almost 90% of the neonatologists surveyed believed the Apgar score to be of little value when assessing hypoxic preterm infants.

The aims of the current study were first, to evaluate the effect of gestational age and respiratory support on interobserver agreement of Apgar scores in preterm infants ≤28 weeks’ gestation compared with term infants and, second, to determine which components demonstrate the least agreement. Our hypothesis was that as gestational age decreases in preterm infants, there would be a decline in interobserver agreement. Our second hypothesis was that as respiratory support decreased, agreement would increase. Our third hypothesis was that muscle tone and grimace, components that depend on the infant’s physiologic maturity, would demonstrate the least agreement.

METHODS

Study Population

Because of their delivery room experience and familiarity with the Apgar score, neonatologists were identified as the most suitable participants for this study. In February 2011, members of the Perinatal Section of the American Academy of Pediatrics received an e-mail describing the study and inviting them to participate in a secure online survey to score 4 cases at 1, 5, and 10 minutes of life. Participants were sent reminder e-mails at 2 and 4 weeks after the initial e-mail and were allotted 2 months for survey completion. Approval for the study was obtained from the University Hospitals Institutional Review Board in Cleveland, Ohio.

Film Clips

Filming occurred between July 2010 and January 2011. Before filming, informed written consent was obtained from 1 parent for each case and verbal consent from the resuscitation team. A Sony digital video camera, mounted on a tripod near the infant’s head or feet, recorded the resuscitation once the infant was placed on the warmer table. Final Cut Express (Apple, Inc, Cupertino, CA) was used to edit selected film clips to remove audio and any personal identifiers of both the infants and delivery room personnel.

Case Descriptions

The survey consisted of 4 delivery room cases that appeared in random order for each participant. However, the individual clips at 1, 5, and 10 minutes of life for each case were not randomized. At 1 minute of life, 5 minutes of life, and for the preterm infants, 10 minutes of life, respondents were provided with the infant’s heart rate and viewed 30 seconds of the delivery room resuscitation. The participant was unable to save answers and return to finish the survey at a later time. Case 1, the study control, depicted a full-term infant (≥38 weeks’ gestation) in room air who was vigorous and breathing spontaneously at 1 and 5 minutes of life. This case was used to establish the expected and acceptable value for observer agreement. In case 2, participants were shown a preterm infant, ~24 weeks’ gestation, presenting with minimal and ineffective respiratory effort, low heart rate, and poor tone. Despite positive pressure ventilation (PPV) with a t-piece resuscitator, the infant became limp (no flexion) with no visible respiratory effort shortly after 1 minute of life. This period of marked deterioration lasted until the infant was intubated. Intubation, at ~8 minutes of life (time interval not shown to viewers), coincided with an improvement in heart rate, activity, and the appearance of some respiratory effort. In case 3, respondents were shown a preterm infant, ~28 weeks’ gestation, who presented with ineffective respiratory effort that quickly improved with initial resuscitation by 1 minute of life. Activity, tone, and heart rate remained stable for the remainder of the resuscitation, which included continuous positive airway pressure (CPAP) with a face mask. In case 4,
survey participants were shown a preterm infant, ~28 weeks’ gestation, who presented with persistent ineffective respiratory effort and was subsequently intubated by 5 minutes of life. It should be noted that during moments of resuscitation when infants were receiving mask ventilation, assessment of grimace (visibility of the eyelids, cheeks, and chin) may have been challenging.

Statistical Methods/Power

Fleiss’ $\kappa$ coefficient was used for multiple raters, a measure of interobserver agreement where a $\kappa$ of 1 equals perfect agreement, a $\kappa$ of 0 equals chance agreement, and a $\kappa$ of −1 equals complete disagreement. Interpretation of $\kappa$ coefficients for this study was based on the standards set forth by Landis and Koch: $\kappa$ < 0 = poor, 0.01 to 0.20 = slight, 0.21 to 0.40 = fair, 0.41 to 0.60 = moderate, 0.61 to 0.80 = substantial, and 0.81 to 1 = almost perfect. The $\kappa$ coefficient presumes that a large number of subjects are evaluated by a small group of expert raters. With a fixed sample size and an acceptable level of $\kappa$ determined to be 0.80 for this study, at least 5 raters were needed to be 95% confident of achieving a “substantial” level of interobserver agreement. Analysis was performed with R program version 2.13.1 (R Foundation for Statistical Computing, Vienna, Austria), and confidence intervals were calculated according to Fleiss’ methodology.

RESULTS

Responses were received from 335 members of the Perinatal Section of the American Academy of Pediatrics. Of the 335 members who responded to the survey, 312 members provided completed scoring for all 4 cases. Seventy-five percent of the respondents acknowledged ≥12 years of delivery room experience.

Figure 1 is a representation of the distribution of total Apgar scores for each of the 4 cases. When compared with the full-term infant, the distribution of total Apgar scores in the preterm cases was wider at 1, 5, and 10 minutes of life.

For case 1 (term infant, room air), the Apgar components heart rate, respiratory effort, grimace, and muscle tone, demonstrated almost perfect agreement (Table 1). However, the $\kappa$ score for color in the full-term infant was low, demonstrating slight agreement at both 1 and 5 minutes of life.

For case 2 (24 weeks’ gestational age, endotracheal positive pressure ventilation (ET PPV)) when the infant displayed ineffective respiratory effort with PPV at 1 minute of life, agreement was fair (Table 1). Fair agreement also occurred after intubation at 10 minutes of life in the face of weak, ineffective respiratory efforts. However, when the infant showed no respiratory effort at 5 minutes, agreement was almost perfect.

$\kappa$ coefficients for grimace and muscle tone also suggest that agreement was related to the clinical condition and activity of the infant. When the infant was limp and exhibiting little, if any, activity, $\kappa$ scores for grimace and muscle tone demonstrated substantial and almost perfect agreement at 1 and 5 minutes of life. However, when the infant’s clinical condition improved and activity returned, agreement for grimace and muscle tone decreased to fair and slight at 10 minutes of life. Of note, $\kappa$ scores for color demonstrated moderate agreement at 1 and 5 minutes only slight agreement at 10 minutes.

For case 3 (28 weeks’ gestational age, CPAP), respiratory effort demonstrated slight agreement when the infant received PPV with a mask at 1 minute of life (Table 1). When the infant received CPAP with a nasal mask, agreement remained slight at 5 minutes of life and improved to fair by 10 minutes of life. Agreement for both grimace and muscle tone was also generally low. Agreement for grimace at 1, 5, and 10 minutes of life was slight whereas muscle tone demonstrated slight agreement at 1 minute of life and fair agreement at 5 and 10 minutes. Color at 1, 5, and 10 minutes of life demonstrated a range of agreement from slight to fair.

For case 4 (28 weeks’ gestational age, ET PPV), respiratory effort demonstrated fair agreement at 1 minute of life when the infant was receiving PPV by mask and slight agreement when intubated at 5 and 10 minutes of life (Table 1). $\kappa$ coefficients for grimace demonstrated fair agreement at 1 minute of life, moderate agreement at 5 minutes of life, and slight agreement at 10 minutes of life. Muscle tone demonstrated slight agreement at 1 and 10 minutes of life and fair agreement at 5 minutes of life. Color demonstrated moderate agreement at 1 minute of life and slight agreement at 5 and 10 minutes of life.

DISCUSSION

This study suggests that interobserver agreement for respiratory effort, grimace, and muscle tone demonstrates considerable disagreement when Apgar scores are assigned to preterm infants ≤28 weeks’ gestation, compared with active term infants in whom agreement is strong for all components measured except color. A marked exception is the strong agreement noted in preterm infants when they are apneic and limp, suggesting that observers agree when the preterm infant’s condition is extremely poor but disagree when some respiratory effort and activity are present. This observation in the extremely sick infant is consistent with Apgar’s own conclusion that “variation is rare in infants with high or low scores.”

$\kappa$

Figure 1
We propose that a significant factor contributing to disagreement in scoring respiratory effort is rooted in the level of respiratory intervention, including CPAP and intubation. With newer methods of noninvasive ventilation being used and recent data suggesting a beneficial role for CPAP over prophylactic intubation in the delivery room,\textsuperscript{17,18} confusion surrounding scoring of respiratory efforts is likely to persist as Apgar scores remain in use during resuscitation. Curiously, while noting that 37% of the infants in her study received PPV by mask or endotracheal tube, Apgar did not speculate on how the type of airway intervention would affect scoring of respiratory effort.\textsuperscript{1} The Committee on Fetus and Newborn previously introduced an “Expanded APGAR score” that considered different modalities of respiratory intervention, but to date this score has not been widely accepted for use.\textsuperscript{19} Alternatively, elimination of the respiratory effort component with a more objective measure of respiratory status such as mode of respiratory support may lead to improved interobserver agreement.

Both grimace and muscle tone exhibited variable agreement in the 3 preterm case scenarios. As in scoring for respiratory effort, agreement was good when respondents judged the infant to be limp but demonstrated disagreement when the infant displayed any signs of activity. The concern that observers do not agree on assessing grimace and muscle tone is not novel. Previous studies have suggested that developmental immaturity influences the score infants receive for these Apgar components.\textsuperscript{20} Apgar’s original paper stated “the usual testing method” for scoring grimace involved “suctioning the oropharynx and nares with a soft rubber catheter,” a practice no longer recommended or routinely performed. Today, face masks and intubation equipment, items often used during resuscitation of preterm infants (including during our study), can create difficulties in assessment of grimace. Even if the mask was removed briefly to assess grimace, this action has the potential in the smallest of preterm infants to cause de-recruitment of alveolar space and hypoxia.

**FIGURE 1**
Distribution of total Apgar scores for each case. Compared with the total Apgar scores in the full-term infant (A [case 1: 38 weeks’ gestation, room air]), the distribution of total Apgar scores for the 3 preterm cases (B [case 2: 24 weeks’ gestation, ET PPV], C [case 3: 28 weeks’ gestation, CPAP], and D [case 4: 28 weeks’ gestation, ET PPV]) was wider at each time point.
Of the 335 neonatologists who participated in the survey, an overwhelming majority (n = 312) provided complete scoring for all 4 cases at each time point. Based on e-mail feedback, a small number of users experienced technical issues that prohibited various clips from appearing correctly. We also speculate that some users may have failed to register their answers when using their mouse to click answer choices.

Finally, although we did not include as an original aim for study the evaluation of color, described by Apgar as “by far the most unsatisfactory sign,”1 we did find significant disagreement in each of the 4 scenarios. This was the only clear point of disagreement in the healthy term infant. Although some of the variation in our study can be attributed to computer monitor settings, O’Donnell et al.12 have previously demonstrated disagreement when Apgar scores are assigned for color.

The study limitations were related mostly to our survey design. A survey length >15 minutes would contribute significantly to participant fatigue. As a result, the online survey was constrained to 4 cases with a limited range of gestational ages. We also reduced the clip duration to 30 seconds. However, one could argue that because these scores are intended to be assigned quickly, 30 seconds would provide ample time for infant assessment. Finally, despite our best efforts to include satisfactory camera angles of the infant, we understand that some participants may have found evaluation of some components difficult. However, adequate visualization of the infant during the resuscitative process is often challenging even when participants are present in the delivery room.

### CONCLUSIONS

This study calls into question 4 of the 5 parameters used when Apgar scores are applied to preterm infants. Score refinements are needed if the score is to continue to provide useful information regarding resuscitative decisions made at birth. An improved score may lead to consistency in assessment among medical care professionals when describing the clinical status of an infant by more accurately reflecting the physiologic immaturity of the preterm infant and by incorporating the latest advances in respiratory intervention.

### ACKNOWLEDGMENTS

We thank members of the Scholarship Oversight Committee; Nancy Cossler, MD, Juliann DiFiore, BSEE, the Perinatal Section of the American Academy of Pediatrics; Eileen Stork, MD; and the physicians who participated in the survey.

### REFERENCES


### TABLE 1: \( \kappa \) Coefficients for All Cases

<table>
<thead>
<tr>
<th>Minute of Life</th>
<th>Description</th>
<th>Heart Rate</th>
<th>Respiratory Effort</th>
<th>Grimace</th>
<th>Muscle Tone</th>
<th>Color</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \kappa )</td>
<td>95% CI</td>
<td>( \kappa )</td>
<td>95% CI</td>
<td>( \kappa )</td>
<td>95% CI</td>
</tr>
<tr>
<td>Case 1</td>
<td>1 (room air) Spontaneous respiratory effort, vigorous</td>
<td>0.97</td>
<td>0.96–0.98</td>
<td>0.94</td>
<td>0.93–0.95</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>5 (room air) Spontaneous respiratory effort, vigorous</td>
<td>1</td>
<td>—</td>
<td>0.91</td>
<td>0.90–0.92</td>
<td>0.90</td>
</tr>
<tr>
<td>Case 2</td>
<td>1 (mask PPV) Ineffective respiratory effort, no flexion</td>
<td>0.94</td>
<td>0.93–0.95</td>
<td>0.93</td>
<td>0.92–0.94</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>5 (mask PPV) No respiratory effort, no flexion</td>
<td>0.97</td>
<td>0.96–0.98</td>
<td>0.87</td>
<td>0.86–0.88</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>10 (ET PPV) Ineffective respiratory effort, some flexion</td>
<td>0.93</td>
<td>0.92–0.94</td>
<td>0.27</td>
<td>0.26–0.28</td>
<td>0.26</td>
</tr>
<tr>
<td>Case 3</td>
<td>1 (mask PPV) Ineffective respiratory effort, some flexion</td>
<td>0.93</td>
<td>0.92–0.94</td>
<td>0.14</td>
<td>0.13–0.15</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>5 (CPAP) Ineffective respiratory effort, well flexed</td>
<td>0.94</td>
<td>0.93–0.95</td>
<td>0.13</td>
<td>0.12–0.14</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>10 (CPAP) Ineffective respiratory effort, well flexed</td>
<td>0.96</td>
<td>0.95–0.97</td>
<td>0.40</td>
<td>0.39–0.41</td>
<td>0.13</td>
</tr>
<tr>
<td>Case 4</td>
<td>1 (mask PPV) Ineffective respiratory effort, some flexion</td>
<td>0.94</td>
<td>0.93–0.95</td>
<td>0.22</td>
<td>0.21–0.23</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>5 (ET PPV) Ineffective respiratory effort, some flexion</td>
<td>0.94</td>
<td>0.93–0.95</td>
<td>0.07</td>
<td>0.06–0.08</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>10 (ET PPV) Ineffective respiratory effort, some flexion</td>
<td>1</td>
<td>—</td>
<td>0.11</td>
<td>0.10–0.12</td>
<td>0.15</td>
</tr>
</tbody>
</table>

\( \kappa \) coefficients: <0 = poor; 0.01–0.20 = slight; 0.21–0.40 = fair; 0.41–0.60 = moderate; 0.61–0.80 = substantial, and 0.81–1 = almost perfect.14 CI, confidence interval.


15. Donner A, Rotondi MA. Sample size requirements for interval estimation of the kappa statistic for interobserver agreement studies with a binary outcome and multiple raters. *Int J Biostat*. 2010;6(1):Article 31


Evaluation of Interobserver Agreement of Apgar Scoring in Preterm Infants
Monuj T. Bashambu, Halana Whitehead, Anna Maria Hibbs, Richard J. Martin and
Monika Bhola

Pediatrics 2012;130;e982; originally published online September 24, 2012;
DOI: 10.1542/peds.2012-0368

Updated Information & Services
including high resolution figures, can be found at:
/content/130/4/e982.full.html

References
This article cites 19 articles, 4 of which can be accessed free at:
/content/130/4/e982.full.html#ref-list-1

Citations
This article has been cited by 3 HighWire-hosted articles:
/content/130/4/e982.full.html#related-urls

Subspecialty Collections
This article, along with others on similar topics, appears in the following collection(s):
Fetus/Newborn Infant
/cgi/collection/fetus:newborn_infant_sub

Permissions & Licensing
Information about reproducing this article in parts (figures, tables) or in its entirety can be found online at:
/site/misc/Permissions.xhtml

Reprints
Information about ordering reprints can be found online:
/site/misc/reprints.xhtml

PEDIATRICS is the official journal of the American Academy of Pediatrics. A monthly publication, it has been published continuously since 1948. PEDIATRICS is owned, published, and trademarked by the American Academy of Pediatrics, 141 Northwest Point Boulevard, Elk Grove Village, Illinois, 60007. Copyright © 2012 by the American Academy of Pediatrics. All rights reserved. Print ISSN: 0031-4005. Online ISSN: 1098-4275.
Evaluation of Interobserver Agreement of Apgar Scoring in Preterm Infants
Monuji T. Bashambu, Halana Whitehead, Anna Maria Hibbs, Richard J. Martin and Monika Bhola

*Pediatrics* 2012;130;e982; originally published online September 24, 2012;
DOI: 10.1542/peds.2012-0368

The online version of this article, along with updated information and services, is located on the World Wide Web at:
/content/130/4/e982.full.html