

Clinicians Can Accurately Assign Apgar Scores to Video Recordings of Simulated Neonatal Resuscitations

Izhak Nadler, MSc;

Helen G. Liley, MBChB, FRACP;

Penelope M. Sanderson, PhD,
FASSA

Introduction: The Apgar score is used to describe the clinical condition of newborns. However, clinicians show low reliability when assigning Apgar scores to video recordings of actual neonatal resuscitations. Simulators provide a controlled environment for recreating and recording resuscitations. Clinicians assigned Apgar scores to such recordings to test the representativeness of simulator and recordings. Study design was guided by Brunswik's probabilistic functionalism.

Method: Judgment analysis methods were used to design 51 recordings of neonatal resuscitation scenarios, simulated with SimNewB (Laerdal, Stavanger, Norway). A step-by-step explanation of the design, preparation, and testing of the recordings is provided.

Analysis: Recorded Apgar scores, calculated from the presentation of clinical signs, were compared against the designed scores. Working independently and without feedback, three experts assigned Apgar scores to confirm that the recordings could be interpreted as intended. Seventeen neonatal resuscitation clinicians scored the recordings in a separate experiment.

Results: Correlations between Apgar scores assigned by the 20 viewers (experts plus clinicians) and recorded Apgar scores were high (0.78–0.91) and significant ($P < 0.01$). Fourteen of the 20 viewers scored the recordings without significant bias. Correlations between viewers' scores and scores of individualized linear models calculated for each viewer were high (0.79–0.97) and significant ($P < 0.01$), indicating systematic judgments.

Conclusions: SimNewB provided a realistic presentation of clinical conditions that was preserved in the recordings. Clinicians could interpret clinical conditions systematically and accurately without feedback or detailed instructions. These methods are applicable to future research about accuracy of clinical assessments in actual and simulated environments.

(*Sim Healthcare* 5:204–212, 2010)

Key Words: Video, Recording, Simulator, Scenario, Policy, Judgment, Apgar, Neonatal resuscitation.

The Apgar score was introduced more than half a century ago to support clinicians' assessments of the condition of newborn infants.¹ The score is computed by summing the contribution of five clinical signs (see Table 1). The original intention was that the score would be calculated 1 minute

after birth to prompt the most appropriate treatment to improve or maintain the newborn's adaptation to extrauterine life. It is now usually also assessed at 5 minutes and then at 5-minute intervals for infants requiring resuscitation, to assess response. The Apgar score has stood the test of time and has become an internationally accepted method of scoring judgments about the physiological status of newborns infant in case notes.^{2–4} Apgar scores are also widely used in clinical audit and research, but there has been surprisingly little research into how accurately they are assigned.

Recently, simulators have been specifically designed for healthcare practitioners to hone neonatal resuscitation skills.⁵ Simulators provide an environment in which relevant clinical variables can be controlled with no risk to patient safety.^{5–7} The SimNewB simulator (Laerdal, Stavanger, Norway) generates clinical signs that are analogous to the components of the Apgar score. Clinicians are expected to assess the signs to decide about appropriate treatment. The contribution of the simulator variables to the Apgar score is presented in Table 1. However, the success of the simulator in use depends on whether practitioners can assess the clinical signs in a reliable manner or can be trained

From the Schools of Information Technology and Electrical Engineering (ITEE) (I.N., P.M.S.), Psychology and Medicine (P.M.S.), The University of Queensland, Brisbane, Queensland; Division of Neonatology, The Mater Mothers' Hospital and Mater Mothers' Research Centre (H.G.L.), Brisbane, Australia.

This paper was written as part of Izhak Nadler's PhD studies at The University of Queensland. During this period, Nadler has held Endeavor IPRS and UQLAS scholarships from The University of Queensland. Nadler's research activities are supported financially in part by grants from the Mater Mothers' Research Centre, Mater Health Services Brisbane, Ltd., Australia, and by an unrestricted donation from Laerdal, Inc.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.simulationinhealthcare.com).

Reprints: Izhak Nadler, School of ITEE, The University of Queensland, Brisbane QLD 4072, Australia (e-mail: itsik@itee.uq.edu.au).

Copyright © 2010 Society for Simulation in Healthcare
DOI: 10.1097/SIH.0b013e3181dcfb22

Table 1. Comparison Between the Contributions of the Clinical Signs When Used for Assigning Apgar Scores to Babies¹ and When Used for Assigning Apgar Scores to the SimNewB Simulator

	Apgar Scores		
	0	1	2
Baby's heart rate	Absent	<100 beats/min	>100 beats/min
Mannequin's heart rate	Absent	<100 beats/min	>100 beats/min
Baby's respiration	Absent	Weak or irregular	Regular breathing
Mannequin's breathing	Absent	<30 breath/min	>30 breath/min
Baby's muscle tone	None	Some flexion	Active movement, good tone
Mannequin's muscle tone	None	Tone	Movement
Baby's reflex	No response to stimulation	Grimace when suctioned	Active withdrawal when suctioned
Mannequin's vocal sounds	None	Weak cry, hiccup, grunting	Strong cry, scream, normal cry
Baby's skin cyanosis	Blue all over	Body pink, extremities blue	No cyanosis, body and extremities pink
Mannequin's simulated oxygen saturation*	<76%	76–82%	83–100%

*During the initial survey of SimNewB capabilities, it was recognized that oxygen saturation readings could not be set to the low levels that were required. This shortcoming was resolved promptly by the simulator developer in collaboration with simulation coordination personnel in the Skills Development Centre.

to do so. Reliability can be determined by measuring how accurately and systematically clinicians assign Apgar scores when compared with Apgar scores calculated algorithmically from clinical signs presented by the simulator.

A useful and an efficient way to make such a comparison is to ask clinicians to assign Apgar scores to recorded scenarios that they view, rather than while taking part in a hands-on simulation. Nonetheless, simulators provide an environment in which, during a short-time frame, relevant clinical variables can be controlled to produce recordings that meet an investigator's requirements with no risk to patient safety.^{5,6,8} Moreover, many variables that could conceivably interfere with the assignment of Apgar scores are eliminated in a recording. Such recordings can be then scored by many observers without needing to spend time in hands-on simulation. If recorded scenarios can be scored reliably then they could be used to support different research and training objectives. However, to date, there has been no test of whether clinicians can reliably assign Apgar scores to video recordings of simulated resuscitations.

Although video recordings of actual clinical settings have been used for auditing purposes and training,^{7,9,10} a recent study¹¹ found that clinicians cannot assign Apgar scores to video recordings of actual neonatal resuscitations with acceptable levels of interobserver reliability, and that observers' scores had little agreement with the original scores assigned by the clinicians who performed the resuscitation. The study¹¹ did not determine the sources for nonreliability, but reasons could include the subjective nature of some components of the Apgar scores, uncertainty as to how to calibrate the score for changes caused by treatment, and difficulty interpreting the video recordings. Because there is no way of knowing what the "true" Apgar scores were for the babies in the recordings, the sources of inaccuracy are hard to discern, whereas in recordings of simulated resuscitation, the true Apgars can be programmed.

To avoid the above potential limitations in our study, the content and the apparatus of recorded scenarios were carefully designed,^{12,13} and the design was guided by solid theoretical principles. The methodology we use was based on a theoretical approach developed by Brunswik¹⁴ and on applications of judgment analysis.¹⁵ Because Brunswik's ideas are

not well known, in the next section we provide some background about his ideas and why we use them in this study.

Brunswik's Theory and Judgment Analysis Background

Brunswik argued that psychologic experiments conducted in laboratories, in which variables are omitted or controlled, create unrealistic situations for participants. Therefore, the results and conclusions of such studies cannot be generalized outside the laboratory. As an alternative, Brunswik^{14,16} introduced the theory of probabilistic functionalism, which, in his view, explains how an organism (subject) interprets cues from its ecology (environment) and makes decisions based on this interpretation. According to Brunswik, experiments—including experiments in laboratories—should have a representative design in which participants are exposed to situations that represent the range and distribution of situations and cues in their natural environment.

Although Brunswik's ideas were controversial,¹⁷ they have increasingly found favor in the human factors community^{18,19} and in other domains.^{20,21} Some of Brunswik's followers worked in the area of judgment analysis and introduced a quantitative model^{22–24} to describe how accurately subjects make judgments about environmental situations.

A linear mathematical model, the so-called "judgment policy," can be used to predict how a clinician will judge a situation. A participant (eg, a clinician in our study) unconsciously, and therefore subjectively, associates a relative importance (weight) to each clinical sign that he or she uses when interpreting the clinical condition of the mannequin. The Apgar score that the clinician assigns reflects this subjective interpretation. To "capture" the policy, the clinician needs to assign Apgar scores to quite a few situations. The values of the cues in each situation need to be known to the researcher. The Apgar score is produced mathematically as the sum of the values of the cues, each multiplied by its relative weight. The values of the weights are computed through a regression that produces the best-fitting linear model for all the assignments made. Once the relative weights are known, the clinician's assignments can be predicted for any given set of Apgar scores. For more details about Brunswik and judgment analysis, see Hammond and Stewart,²⁰ Cooksey,¹⁵

Table 2. Terminology for Different Sets of Apgar Scores

Term	Definition
Original Apgar scores	Apgar scores for the simulated cases, at the initial design phase
Recorded Apgar scores	Apgar scores calculated from the clinical signs presented in the video recordings
Judged Apgar scores	Apgar scores assigned by a viewer for each of the recordings
Policy Apgar scores	Apgar scores calculated from the linear judgment policy of a viewer

Karelaia and Hogarth,²⁵ Kirlik,^{18,19} and Wigton.²¹ For studies in healthcare that used these concepts, see the studies by Wigton et al,²⁶ Beckstead and Stamp,²⁷ and Thompson et al.²⁸

Relevance

Brunswik's theory was a natural choice for this study because we wanted to evaluate how accurately neonatal resuscitation practitioners can assign Apgar scores to a simulator. Results of such an evaluation can then be used to determine the representativeness of the scenarios and to demonstrate whether the simulator can be used to prepare scenarios that are relevant to neonatal resuscitation practice.

Objectives

In this article, we describe the design, production, and initial validation of a compilation of recordings of simulated neonatal resuscitations informed by the above approach. After the initial validation, we used the compilation in a study in which clinicians were tasked with assigning Apgar scores to each recording. We wanted to determine the feasibility of the overall approach, to measure the accuracy of Apgar scoring in this setting, and to determine the representativeness of the recordings and scenarios developed.

MATERIALS AND METHODS

Table 2 shows the terminology we used to describe different sets of Apgar scores developed during the production and the validation of the recordings. The entire process of designing, producing, and analyzing the recordings is presented in Table 3 and is described in the sections that follow. It should

be noted that when designing recordings for purposes different from ours, other factors may need to be considered.

Initial Design of the Scenarios

Number of Scenarios

Following judgment analysis guidelines¹⁵ for obtaining each viewer's judgment policy, and in order to expose viewers to the full range of relevant situations, the number of recordings should be between 5 and 10 times the number of clinical signs that are used to interpret the situation (ie, for the five-component Apgar score, 25–50 recordings). Therefore, we decided to develop about 50 recordings.

Length of Scenarios

The length of each recording was designed according to two principles. First, the compilation needed to be short enough to allow a viewer to maintain concentration while studying all the scenarios. Second, each recording needed to be long enough to tell a story and engage the viewer, and to permit stable clinical assessment of the mannequin. We concluded that each recording should last about 2 minutes, allowing viewers who are used to the 30-second assessment intervals recommended in current resuscitation guidelines²⁹ ample time for assessments while viewing each recording. We estimated that a total viewing time of about 2 hours for the 50 recordings would let the viewers comfortably score the entire compilation while maintaining their concentration. The final length of recordings in the compilation was designed to vary between 1.45 and 2.15 minutes.

Overall Apgar Scores

The first step in designing the clinical characteristics for the scenarios was done by defining Apgar scores at the start, middle, and end of each scenario. To conform to Brunswik's concept of representative design¹⁴ in which subjects should be exposed to a range of situations similar to their daily routine, we wanted the clinical characteristics to be typical of actual neonatal resuscitation cases. We scanned computerized birth records for all babies born in the Mater Mothers' Hospital in Brisbane, Australia, in the first 6 months of 2008. For those babies who had an Apgar score of 5 or less at 1 minute, a midwifery research assistant undertook a clinical record review and noted the Apgar score at 1, 5, and 10 minutes. Apgar scores of 5 or below at 1 minute are

Table 3. Summary of the Processes of Developing and Testing the Recordings

Phase	Considerations and References	Process
1. Initial design of scenarios	1a. Judgment Analysis guidelines ¹⁵ 1b. Clinical guidelines ²⁵ and total viewing time limits 1c. Representative design ¹⁵ indicates that patterns of Apgar scores be drawn from patient charts	1a. Define recommended number of scenarios 1b. Define length of scenarios 1c. Define overall Apgar scores
2. Detailed design of scenarios	2a. Relationship between clinical signs, ⁴ Table 4 2b. Apgar score calculation ¹ , Table 1 2c. Scenarios reviewed for face validity by experts	2a. Define contribution of clinical signs to the overall Apgar score 2b. Define specific values of clinical signs 2c. Modifications following face validity review
3. Production of scenarios and recordings	3a. N/A 3b. Medical procedures provided by medical team	3a. Program scenarios in the SimNewB™ 3b. Record scenarios in a simulation center with clinician actors
4. Rating of the compilation and use	4a. Kolmogorov-Smirnov test, t-test, Pearson product-moment correlation, Table 5 4b. t-test, Pearson product-moment correlation, Table 6 4c. t-test, Pearson product-moment correlation, Table 7	4a. Test representativeness of recordings 4b. Validation of recordings by experts 4c. Clinicians' assignment of Apgar scores in an experiment

usually assigned to infants who are receiving, or who are recognized to need, at least some resuscitation maneuvers.

By interpolation from the obtained scores, we estimated the distributions of the Apgar scores to which clinicians were exposed during the resuscitation period and we could also estimate patterns of changes in the Apgar scores during this period. These estimations were used to define the distribution of the Apgar scores at the start, middle, and end of each recording (“Original Apgar Scores”). This process was used to define the Apgar scores for 42 scenarios.

Apgar scores for a further nine scenarios were defined in a somewhat more artificial manner. Our survey did not yield patient charts indicating actual deterioration patterns. Although deterioration in the condition of the newborn during resuscitation is not much analyzed in the literature, clinical experience suggests that it is common. To define Apgar scores for the start, middle, and end of these nine scenarios, we randomly picked nine of the cases that had been previously designed and modified them to present deterioration patterns.

Detailed Design of the Scenarios

To define the values of the clinical signs for the scenarios, we needed to know the contribution of each sign to the total Apgar score at the start, middle, and end points of each scenario. This breakdown was not available in most patient charts, and only total scores were available. Therefore, we “reverse engineered” what the corresponding clinical signs could have been, using the following procedure.

Probable Contributions of Clinical Signs to Apgar Scores

The physiological responses that comprise the Apgar score are not independent. During deterioration and recovery, they will often follow a typical sequence.⁴ For example, heart rate and respiratory effort will usually deteriorate after, and recover before, color and tone. By consulting a neonatal resuscitation medical expert, we produced the probable contribution of each sign to the various total Apgar scores. These contributions are presented in Table 4.

Calculating the Contributions of Clinical Signs to Apgar Scores

We explain this process by the following example. Consider a case in which the intended Apgar score is 5 and we are considering the contribution of heart rate to this Apgar score. First, a random number between 0 and 1.0 is generated. Second, referring to Table 4 for an overall Apgar score of 5, the contribution of the heart rate to the total Apgar score will be 2 if the random number is between 0.8 and 1.0, the contribution will be 1 if the random number is 0.8 or below, and there is 0.0 likelihood of heart rate contributing 0 if the Apgar score is to be 5. This process of defining the contribution for each of the five clinical signs was repeated for all the Original Apgar Scores.

Mapping Contributions to Clinical Signs

Once the contribution of each clinical sign was defined, we used Table 1 to link the contributions with actual values of the clinical signs. In cases in which the contribution indicates a range for the possible values (as with the heart rate, respiration rate, and oxyhemoglobin saturation) or several options (as with the vocal sound and muscle tone), the specific value was randomly selected. The change in the value of each clinical sign was designed to start at a randomly selected moment within the first 15% of the scenario length, reach its middle value at a randomly selected moment within 15% of the middle of the scenario, and reach its end value at a randomly selected moment within the last 15% of the scenario length.

Face Validity

The specifications for each simulated case were reviewed by medical experts. In response to the experts’ comments, slight adjustments were made in a few cases in which the randomization process had produced a combination of values that seemed physiologically implausible. The values of the clinical signs were used for programming the scenarios into the simulator.

Production of the Scenarios and the Recordings Programming the Simulator

The Laerdal SimNewB simulator was chosen for this study because it was specifically developed to support neonatal re-

Table 4. The Contributions of Different Clinical Signs to the Apgar Scores

Total Apgar	Probable Contributions of Clinical Signs to the Apgar Score														
	Heart Rate			Breathing			Blood Saturation			Muscle Tone			Vocal Sound		
	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0
1	0	1	0	1	0	0	1	0	0	1	0	0	1	0	0
2	0	1	0	0	1	0	1	0	0	1	0	0	1	0	0
3	0	1	0	0	1	0	0.65	0.35	0	0.65	0.35	0	0.7	0.3	0
4	0	1	0	0	1	0	0.3	0.7	0	0.3	0.7	0	0.4	0.6	0
5	0	0.8	0.2	0	0.9	0.1	0.1	0.9	0	0.1	0.9	0	0.1	0.9	0
6	0	0.6	0.4	0	0.7	0.3	0	0.9	0.1	0	0.9	0.1	0	0.9	0.1
7	0	0.35	0.65	0	0.45	0.55	0	0.7	0.3	0	0.7	0.3	0	0.8	0.2
8	0	0.2	0.8	0	0.3	0.7	0	0.5	0.5	0	0.5	0.5	0	0.5	0.5
9	0	0.1	0.9	0	0.1	0.9	0	0.25	0.75	0	0.25	0.75	0	0.3	0.7
10	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1

Each clinical sign can contribute 0, 1, or 2 points to the Apgar score, depending on the neonate’s state. Values in the body of the table are the probability that each clinical sign will be contributing 0, 1, or 2 points to the Apgar score shown at the left of the row.



Figure 1. Snapshot from a recording of a scenario with the SimNewB simulator, showing the heart rate and blood saturation superimposed over the recording.

suscitation practitioners and researchers. A detailed survey of the simulator, exploring its capabilities and limitations, was performed as part of the design phase. A summary of our findings is presented herein.

Heart rate (manifesting as heart sounds audible with a stethoscope and pulsations palpable in the umbilical cord stump and brachial area) and breathing (breath sounds audible with a stethoscope and visible as chest wall movement) are the most realistically simulated clinical signs. Expiratory “grunting,” a typical sign of respiratory distress in newborns, and crying can also be simulated. A patient monitor can be used to display heart rate and oxyhemoglobin saturation readings. Skin color over the face can be changed from bluish to pink, but this capability was difficult to see clearly in the recordings. Alterations in muscle tone and movements can also be simulated. Representing “reflex irritability” is more difficult, although some limb movement or vocalization can be used.

The clinical signs and their changes over time were programmed into the SimNewB simulator, mainly by using its object-oriented interface. The run-time algorithm of the simulator introduces slight variability to the programmed values. The variability is mainly intended to produce realism, because human heart rate and oxygen saturation levels usually vary around a constant level or trend.

Video Recording of the Scenarios

The scenarios were performed and recorded in Queensland Health’s Skills Development Centre at Herston in Brisbane. Figure 1 represents the video setting of the recorded scenarios. The patient monitor was recorded as a “picture in picture” superimposed over the video recording of the mannequin. A spotlight was used to create shadows on the chest of the mannequin to accentuate chest movements in the playback. Figure 2 presents the apparatus in the recording environment.

For recording the heart and the lung sounds, a microphone was installed under the “skin” of the mannequin, and the sounds were recorded on one of the two stereo-audio channels. The patient monitor alarms and the vocal sounds generated by the mannequin were recorded on both chan-



Figure 2. Recording apparatus showing SimNewB in the resuscitation cot, video camera on the tripod, a spotlight to the right of the cot, and the patient monitor to the left.

nels. This method made it possible to play the heart and lung sounds into headsets or into modified stethoscopes to increase the realism of the presentation. To prevent situations in which a mannequin ventilated with an endotracheal tube still produced vocal sounds, we recorded the vocal sounds off-line and then edited them into the recordings.

During each recorded scenario, a resuscitation team provided the mannequin with simulated medical treatment. Only the hands and the forearms of team members were captured in the field of view of the camera. We specifically directed the team to perform some activities that were obviously insufficient or inappropriate to the situation in order to prevent viewers from assuming that the condition of the baby would usually improve on the basis that competent resuscitation was being provided. We also instructed the team members to avoid verbal communication during recording.

Viewers of the compilation who later assigned Apgar scores to the recordings were advised about these constraints and were directed to base their clinical assessment solely on the clinical signs and not on the quality or type of clinical procedures. It was also decided that data such as gestation age and presenting medical history would be concealed from later viewers, because it could bias the assignment of the Apgar scores, as happens in clinical practice.³

To illustrate the final product of the development, an example taken from one of the recordings is provided in Video, Supplemental Digital Content 1, <http://links.lww.com/SIH/A10>.

Rating of the Compilation

Rating by Experts

The compilation was reviewed and independently scored by three neonatal resuscitation experts—two consultants,

plus one neonatal nurse who is also a neonatal resuscitation instructor. After viewing each recording, the experts were asked to assign an Apgar score that reflected the condition of the mannequin at the end of the recording. As noted earlier, the experts were asked to base their judgment on the clinical signs and to ignore the appropriateness of the care provided by the resuscitation team. The experts were not exposed to the specific mapping that linked the simulator's clinical sign values to the Apgar score, and they were not given feedback about the correctness of their scoring.

Rating of the Compilation During a Judgment Accuracy Experiment

In the experiment for which the compilation was prepared, each recording was projected on a wide screen, allowing groups of five or six participants at a time to comfortably view the scenario. Participants used Cardionics Hertman simulated stethoscopes (Webster, TX) to listen to the heart and breath sounds of the mannequin. Other vocal sounds and alarms were displayed with a free-field speaker as well as on the simulated stethoscope. After concerns were expressed by the experts about the length of the rating process, the number of recordings was reduced, and we used only 30 recordings for the analysis. Participants were given rest breaks after every 10 recordings and did not receive any feedback about the accuracy of their scoring.

The experiment was approved by the Human Research Ethics Committees of the Mater Mothers' Hospital, Queensland Health, and by The University of Queensland.

Statistical Analysis

Testing the Representative Design of the Recordings

To make sure that the effects of representative design were still in place by the end of the design and production processes, we analyzed the recorded scenarios to ensure that they had reproduced the Original Apgar Scores. Apgar scores for the mannequin were calculated for the start, middle, and end of each recording ("Recorded Apgar Scores"). The calculation was made by using Table 1 and the values of the clinical signs. There was no need for subjective interpretation in the process; some values (eg, heart rate) were taken from the recorded patient monitor and other values were obtained from the log files that were generated by the simulator while executing the scenarios that we recorded. Kolmogorov-Smirnov tests (Statistica version 8) and Pearson product-moment correlations were used to verify that the Recorded Apgar Scores and the Original Apgar Scores did not differ significantly from each other.

Analysis of the Compilation by Experts

We calculated Pearson product-moment correlation coefficients to determine whether the unaided experts could produce Apgar scores ("Judged Apgar Scores") that were closely positively correlated with the Recorded Apgar Scores taken at the end of the recordings. In addition, *t* tests for significant differences between the means of the two sets of Apgar scores were calculated to determine whether the experts' Judged Apgar Scores were biased with respect to the Recorded Apgar Scores.

We also tested whether the experts could achieve a systematic interpretation of the clinical signs. A judgment policy was calculated for each expert. By feeding the linear model with the values of the clinical signs observable at the end of each recording, a set of estimated Apgar scores ("Policy Apgar Scores") was calculated. A high correlation between an expert's Judged Apgar Scores and his or her Policy Apgar Scores indicates that the expert has a systematic and reproducible policy for combining the clinical evidence and assigning Apgar scores.

Scoring of the Compilation by Neonatal Clinicians in Judgment Accuracy Experiment

Having confirmed that the experts could reliably assign Apgar scores to the compilation, we proceeded to a judgment accuracy experiment. Participants in this experiment were 17 neonatal practitioners who had varying levels of training and experience. The participants included two neonatologists, three neonatal fellows, four pediatric registrars equivalent to postgraduate year-1 to -3 residents, and eight neonatal nurses. The participants viewed and assigned Apgar scores to 30 of the recorded scenarios. Pearson product-moment correlations were calculated between the Judged Apgar Scores of the 17 clinicians and the Recorded Apgar Scores calculated directly from the recordings. Separate *t* tests were used to identify any significant bias in the judgments of each clinician. Correlations between the Judged Apgar Scores and the Policy Apgar Score were calculated to test how systematic and reproducible the clinicians' judgments were.

RESULTS

Validation of Representative Design of the Scenarios

A comparison of the distributions of the Original Apgar Scores and the Recorded Apgar Scores suggests that the recordings matched the original design as intended (Table 5). Specifically, the nonsignificant results for the Kolmogorov-Smirnov test and the significant positive correlation between

Table 5. Comparing the Original Apgar Scores (At the Initial Design Phase) With the Recorded Apgar Scores (as Calculated From the Recorded Scenarios)

	Max Neg. Differ.	Max Pos. Differ.	Mean (SD) of Original Apgar Score	Mean (SD) of Recorded Apgar Score	<i>P</i> Level of KS++ Test	Pearson Correlation Between Scores (<i>r</i>)
Apgar start+	-0.06	0.06	3.49 (2.20)	3.57 (2.43)	>0.10	0.87*
Apgar mid+	-0.04	0.04	3.84 (2.29)	3.80 (2.34)	>0.10	0.88*
Apgar end+	0.02	0.06	4.25 (2.43)	3.98 (2.31)	>0.10	0.85*

+The comparison was done for the Apgar scores at the start, middle, and end of each scenario.

*Significant at *P* < 0.01.

++KS indicates Kolmogorov-Smirnov test that compared the distributions.

Table 6. Results for Three Experts Who Viewed and Scored the Recordings

	Judged Apgar Scores Mean (SD)	Policy Apgar Scores Mean (SD)	Recorded Apgar Scores vs. Judged Apgar Scores		Judged Apgar Scores vs. Policy Apgar Scores Correlation (<i>r</i>)
			Correlation (<i>r</i>)	<i>P</i> (<i>T</i> < = <i>t</i>)	
Expert 1	5.92 (2.35)	5.92 (2.10)	0.79*	<.001	0.89*
Expert 2	3.45 (2.52)	3.45 (2.21)	0.80*	N/A	0.88*
Expert 3	3.51 (2.54)	3.51 (2.00)	0.73*	N/A	0.79*

The Judged Apgar scores (scores assigned by the experts) are compared with the Recorded Apgar scores (as calculated from the recorded scenarios) and with the Policy Apgar Scores (calculated for each participant, via a linear regression model reflecting relative importance of each clinical sign during Apgar score assignment).

*Significant at $P < 0.01$.

the two sets of scores suggest that the recordings matched the original design as intended.

Validation of Scenarios by Experts

The results for the experts' Judged Apgar Scores and Policy Apgar Scores are presented in Table 6 and are discussed below.

Judged Apgar Scores Versus Recorded Apgar Scores

The mean and standard deviation of the Recorded Apgar Scores at the end of the recordings were 3.98 and 2.31, respectively (Table 5). The means and standard deviations in Table 6 indicate that expert 1 observed and responded appropriately to the variability in the recorded scenarios, but made ratings that were 2.3 Apgar units above the Recorded Apgar Scores. A *t* test indicated that the ratings of expert 1 were significantly higher ($P < 0.001$) than the recorded scores. The other two experts also responded appropriately to the variability with no significant bias in their judgments.

Fit of Judgments to Policy

The significant positive correlations between experts' Judged Apgar Scores and their Policy Apgar Scores suggest

that the experts were systematic in their interpretation of the clinical signs when assigning the Apgar scores.

Clinicians' Assignments of Apgar Scores to Recordings in Judgment Accuracy Experiment

There was a significant positive correlation between the judged and the recorded Apgar scores for each of the 17 clinicians ($P < 0.01$; Table 7). The *t* tests between the means of the two sets of scores for each clinician indicated that five of the 17 clinicians—all of them nurses—assigned Apgar scores that were significantly higher than the recorded scores. The bias of three nurses was between +0.6 and +0.8 Apgar scores and that of the other two nurses was between +1.05 and +1.15 Apgar scores. An ANOVA comparing the Apgar scores assigned by doctors ($n = 9$) versus nurses ($n = 8$) indicated that as a group the nurses assigned significantly higher Apgar scores than did doctors ($P < 0.001$), despite judging exactly the same recordings.

Correlations between the 17 clinicians' Judged Apgar Scores and their Policy Apgar Scores were all positive and highly significant ($P < 0.01$), indicating that all clinicians

Table 7. Results for 17 Clinicians in a Judgment Accuracy Experiment Who Viewed and Scored the Recordings

Clinician	Judged Apgar Scores Mean (SD)	Policy Apgar Scores Mean (SD)	Recorded Apgar Scores vs. Judged Apgar Scores		Judged Apgar Scores vs. Policy Apgar Scores Correlation (<i>r</i>)
			Correlation (<i>r</i>)	<i>P</i> (<i>T</i> < = <i>t</i>)	
1 (Doctor)	3.83 (3.69)	3.83 (3.21)	0.83*	N/A	0.87*
2 (Doctor)	4.53 (3.28)	4.53 (3.02)	0.91*	N/A	0.92*
3 (Doctor)	4.23 (2.88)	4.23 (2.62)	0.87*	N/A	0.91*
4 (Nurse)	4.90 (3.56)	4.89 (3.29)	0.91*	<0.05	0.93*
5 (Nurse)	4.36 (3.54)	4.36 (3.16)	0.88*	N/A	0.89*
6 (Doctor)	3.96 (3.28)	3.96 (2.96)	0.88*	N/A	0.90*
7 (Doctor)	4.40 (2.51)	4.39 (2.11)	0.84*	N/A	0.84*
8 (Doctor)	4.16 (3.01)	4.16 (2.82)	0.91*	N/A	0.94*
9 (Nurse)	5.03 (2.47)	5.03 (2.27)	0.91*	<0.05	0.92*
10 (Nurse)	5.33 (2.21)	5.33 (1.97)	0.84*	<0.05	0.89*
11 (Nurse)	5.40 (2.47)	5.39 (2.25)	0.85*	<0.05	0.91*
12 (Doctor)	3.90 (2.95)	3.89 (2.61)	0.91*	N/A	0.89*
13 (Doctor)	4.00 (3.00)	3.99 (2.72)	0.87*	N/A	0.91*
14 (Doctor)	3.96 (2.83)	3.96 (2.54)	0.87*	N/A	0.90*
15 (Nurse)	4.43 (2.54)	4.43 (2.30)	0.85*	N/A	0.91*
16 (Nurse)	4.90 (2.52)	4.89 (2.19)	0.78*	<0.05	0.87*
17 (Nurse)	4.33 (2.24)	4.33 (2.17)	0.87*	N/A	0.97*

The Judged Apgar scores (assigned by the clinicians) are compared with the Recorded Apgar scores (as calculated from the recorded scenarios) and with the Policy Apgar Scores (calculated for each clinician, via a linear regression model reflecting relative importance of each clinical sign during Apgar score assignment).

*Significant at $P < 0.01$.

had a highly systematic interpretation of the clinical signs.

DISCUSSION

When properly used, simulators provide a controlled environment for producing diverse clinical situations that can be video recorded and later used for various purposes.⁸ For example, video recordings have proved to be an excellent training tool,³⁰ but the specific objective—research or training—must inform the design of the recordings. We demonstrated that Brunswik’s theoretical framework of representative design can be merged with requirements imposed by the future use of a compilation of recordings to guide the design of scenarios and serve further studies that investigate the accuracy of judgments in other clinical settings.²¹ As noted earlier, Brunswik’s framework is used for design and evaluation in many other domains. The healthcare simulation community may benefit from incorporating some of Brunswik’s ideas into simulator studies and applications and from collaborating with the research communities that use such ideas.

In this study, all viewers were sensitive to the variability in the clinical situations, with the result that their judgments were highly positively correlated with the clinical situations represented by the simulator.

No doctor (in either the expert or clinician group) showed significant bias between the mean of their Apgar judgments and the mean of the Apgar scores represented by the simulator in the recordings. In contrast, six of the nine nurses (one in the expert group and five in the clinician group) showed a significant bias toward making Apgar judgments that were higher than the scores represented by the simulator in the recordings. The judgment policies of the six nurses assigning higher scores did not indicate any systematic reason for this finding.

The high correlations and the accuracy of the scores that most clinicians assigned suggest that there was a good connection between the task of rating the recordings and clinicians’ daily practice in the hospital, even though they received no feedback about their accuracy during the task and had no prior experience of viewing recordings of simulated neonatal resuscitation events. Such an outcome indicates that the scenarios we created provided a relevant representation of clinical events and that the recordings with SimNewB managed to preserve this relevance. We conclude that the process of assigning Apgar scores was not compromised by physical differences between how clinical signs present in the mannequin versus in a human neonate and that SimNewB provided an environment in which clinical assessments can be performed accurately as a part of training or research.

For clinicians who demonstrated significant bias, a deeper analysis of their judgment policies could elucidate the source or sources of bias. Training combined with feedback could potentially resolve these biases.

In a previous study¹¹ that reported low-interobserver reliability when clinicians assigned Apgar scores to video recordings of actual resuscitations, the clinicians participating in that study did not reproduce the Apgar scores assigned by

the clinicians who performed the actual resuscitations. In contrast, in our study which used recordings of simulated resuscitations, we found good reliability. Possible reasons are as follows.

First, in the simulated environment, the accuracy of clinical assessments are tested against objective data, which makes it possible to identify the accuracy of each clinician over a range of recordings. In recordings of actual settings, there are no “correct” scores, and the accuracy of individuals cannot be computed.

Second, we instructed our viewers to ignore the clinical treatment when assigning the Apgar scores. If such instructions are not indicated up front, there is a possibility that medical activities may be used as indications of the illness severity of the baby. Some other indicators that are not part of the Apgar score calculation, but that exist only in recordings of actual resuscitation, may contribute to the variability of the scores (eg, a baby’s size).

Third, some variability may emerge from the complicated resuscitation environment in which, for example, color and chest wall movements (indicating respiration efforts) cannot be systematically captured and presented. Such constraints are much fewer in the simulated environment.

Overall, researchers wanting to prepare a compilation of actual resuscitation events may benefit from following key aspects of the methodology we have presented for preparing and testing a compilation of recorded events. Such an approach may help researchers pinpoint sources of variability in how clinicians assign Apgar scores to actual recordings.

From a broader perspective, the processes described in this article could guide the production of video recordings intended to support other kinds of investigation. However, any application of the processes described here, although supported by a robust theoretical framework, must take into account the intended use of the recordings and must be tempered by practical concerns. These steps are essential for producing any compilation of recordings that is representative of the clinical phenomena of interest.

When clinicians without feedback or training can accurately and systematically assess the clinical condition of a mannequin by observing recorded scenarios, it is reasonable to conclude that the simulator can generate a realistic representation of clinical signs that are used in actual practice for clinical assessments. The simulator can be used with greater confidence not only for performance studies but also for training, procedure evaluation, and for evaluating new information tools and cognitive aids.

ACKNOWLEDGMENTS

The authors thank participation and kind support from management and staff of the Queensland Healthcare Skills Development Centre and of the Mater Mothers’ Hospital. Most importantly, they acknowledge the assistance given by Daniel Host to the preparation and the conduct of this study. They also thank Dr. Neil Finer for advice and assistance.

REFERENCES

1. Apgar V. A proposal for a new method of evaluation of the newborn infant. *Curr Res Anesth Analg* 1953;32:260–267.

2. Papile LA. The Apgar score in the 21st century. *N Engl J Med* 2001;344: 519–520.
3. The Apgar score. *Pediatrics* 2006;117:1444–1447.
4. Pinheiro JMB. The Apgar cycle: a new view of a familiar scoring system. *Arch Dis Child Fetal Neonatal Ed* 2009;94:F70–F72.
5. Assar D, Chamberlain D, Colquhoun M, et al. Randomised controlled trials of staged teaching for basic life support 1. Skill acquisition at bronze stage. *Resuscitation* 2000;45:7–15.
6. Takiguchi S, Sekimoto M, Yasui M, et al. Cyber visual training as a new method for the mastery of endoscopic surgery. *Surg Endosc* 2005; 19:1204–1210.
7. Carbine DN, Finer NN, Knodel E, Rich W. Video Recording as a means of evaluating neonatal resuscitation performance. *Pediatrics* 2000;106:654–658.
8. Stefanidis D, Korndorffer JR, Heniford BT, Scott DJ. Limited feedback and video tutorials optimize learning and resource utilization during laparoscopic simulator training. *Surgery* 2007;142:202–206.
9. Weinger MB, Gonzales DC, Slagle J, Syced M. Video capture of clinical care to enhance patient safety. *Qual Saf Health Care* 2004;13: 136–144.
10. Scherer LA, Chang MC, Meredith JW, Battistella FD. Videotape review leads to rapid and sustained learning. *Am J Surg* 2003;185:516–520.
11. O'Donnell CPF, Kamlin COF, Davis PG, Carlin JB, Morley CJ. Interobserver variability of the 5-minute Apgar score. *J Pediatr* 2006;149:486–489.
12. Devitt JH. Testing the raters: inter-rater reliability of standardized anaesthesia simulator performance. *Can J Anaesth* 1997;44:924–928.
13. Morgan PJ, Cleave-Hogg D. Evaluation of medical students' performance using the anaesthesia simulator. *Med Educ* 2000;34:42–45.
14. Brunswik E. *Perception and the Representative Design of Psychological Experiments*. 2nd ed. Berkeley, CA: University of California; 1956.
15. Cooksey RW. *Judgment Analysis: Theory, Methods and Applications*. San Diego, CA: Academic Press; 1996.
16. Brunswik E. *The Conceptual Framework of Psychology*. Chicago, IL: University of Chicago Press; 1952.
17. Hammond KR, Stewart TR. Introduction. In: Hammond KR, Stewart TR, eds. *The Essential Brunswik: Beginnings, Explications, Applications*. New York: Oxford University Press; 2001:540.
18. Kirlik A. Brunswikian resources for event-perception research. *Perception* 2009;38:376–398.
19. Kirlik A. Adaptive Perspective on Human-Technology Interaction: Methods and Models for Cognitive Engineering and Human-Computer Interaction. New York: Oxford University Press; 2006.
20. Hammond KR, Stewart TR. *The Essential Brunswik: Beginnings, Explications, Applications*. New York: Oxford University Press; 2001.
21. Wigton RS. What do the theories of Egon Brunswik have to say to medical education? *Adv Health Sci Educ Theory Pract* 2008;13:109–121.
22. Holzworth RJ. Judgment analysis. In: Hammond KR, Stewart TR, eds. *The Essential Brunswik: Beginnings, Explications, Applications*. New York: Oxford University Press; 2001:324–327.
23. Tucker LR. A suggested alternative formulation in the developments by Hursch, Hammond, and Hursch, and by Hammond, Hursch, and Todd. *Psychol Rev* 1964;71:528–530.
24. Hursch CJ, Hammond KR, Hursch JL. Some methodological considerations in multiple-cue-probability studies. *Psychol Rev* 1964; 71:42–60.
25. Karelaia N, Hogarth RM. Determinants of linear judgment: A meta-analysis of lens model studies. *Psychol Bull* 2008;134:404–426.
26. Wigton RS, Darr CA, Corbett KK, Nickol DR, Gonzales R. How do community practitioners decide whether to prescribe antibiotics for acute respiratory tract infections? *J Gen Intern Med* 2008;23:1615–1620.
27. Beckstead JW, Stamp KD. Understanding how nurse practitioners estimate patients' risk for coronary heart disease: a judgment analysis. 2007;60:436–446.
28. Thompson C, Bucknall T, Estabrookes CA, et al. Nurses' critical event risk assessments: a judgement analysis. *J Clin Nurs* 2009;18:601–612.
29. Kattwinkel J, Short J. *Textbook of Neonatal Resuscitation*. 5th ed. Elk Grove Village, IL: American Academy of Pediatrics; 2006.
30. Guise JM, Deering SH, Kanki BG, et al. Validation of a tool to measure and promote clinical teamwork. empirical investigations. *Simul Healthc* 2008;3:217–233.