

# The Accuracy of Clinical Assessments as a Measure for Teamwork Effectiveness

Izhak Nadler, MSc;

Penelope M. Sanderson, PhD,  
FASSA;

Helen G. Liley, MBChB, FRACP

**Introduction:** Team training in healthcare is usually evaluated by observers who either score trainees' behaviors, social skills, and cognitive skills during simulation or measure changes in the clinical state of a mannequin. Both methods have shortcomings that limit their usefulness. We propose Brunswik's probabilistic functionalism and the Accuracy Score (AS), a measure emerging from judgment analysis, as elements of a complementary approach that could increase the objectivity of team training evaluation. We report an initial investigation.

**Method:** Three groups of neonatal clinicians participated in a resuscitation experiment involving three different training interventions. During the experiment, at various phases, the participants were required to assign an Apgar score to a mannequin.

**Analysis:** The AS was used to test how accurately the clinicians assigned Apgar scores to the mannequin across different levels of task demand, training content, and training delivery method.

**Results:** The AS was lower when task demand increased ( $P < 0.01$ ). The AS was higher after teamwork training than after clinical training ( $P < 0.05$ ) and better after hands-on teamwork training than after lecture-based teamwork training ( $P < 0.05$ ).

**Conclusions:** Because it is simple and objective, the AS may complement existing measures for team training evaluation. Future studies are required in which the AS is tested with a larger number of trainees, in longitudinal experiments, across different training areas, and is compared with previously validated team performance measures.

(*Sim Healthcare* 6:260–268, 2011)

**Key Words:** Neonatal resuscitation, Accuracy Score, Judgment analysis, NRP, CRM, Brunswik.

Team training in healthcare is aimed at turning a heterogeneous group of expert clinicians into a team that can perform multiple coordinated activities under stressful conditions.<sup>1,2</sup> There is no single measure that can be used to score team coordination, so previous authors have recommended that the evaluation of team training effectiveness should use several measures that reflect various aspects of team performance.<sup>3,4</sup>

A common approach is to score the behaviors, social skills, and cognitive skills shown by team members during simulation.<sup>4,5</sup> Typically, observers assign scores either while observing the trainees in simulated clinical performance or afterward while viewing recordings of these events.<sup>2</sup> To reduce subjectivity, observers need to be specifically trained for the scoring task.<sup>5–7</sup>

A complementary approach to evaluation involves scoring performance outcomes, such as the eventual clinical state of the mannequin.<sup>2</sup> Outcome measures may be objective, but they do not explain why one team performed better or worse than another.<sup>4</sup> Moreover, in some cases, outcome measures reflect the performance of individual members and not necessarily the performance of the team.

We propose an additional, objective measure of team effectiveness that assesses a different aspect of team function. The measure is based on Brunswik's theory of probabilistic functionalism, described later, and judgment analysis techniques, and it scores the accuracy of clinical assessments made by team members. Brunswik's theory describes how individuals make assessments and act in demanding tasks, suggesting that measuring accuracy of clinical assessment may be a valid and objective way of measuring how well clinical teams perform in diverse scenarios. Furthermore, if team training makes team members' clinical assessment more accurate, then the interventions performed by the team are probably more clinically apt and could result in better patient outcomes.

## PROBABILISTIC FUNCTIONALISM AND JUDGMENT ANALYSIS

Brunswik's theory of probabilistic functionalism<sup>8,9</sup> describes how individuals assess an environmental situation by attending to cues related to the situation and then act on their

From the School of ITEE (I.N.), Schools of ITEE, Psychology and Medicine (P.M.S.), and The Mater Mothers' Hospital and School of Medicine (H.G.L.), The University of Queensland, Brisbane, Queensland, Australia.

Supported by the Mater Mothers' Research Centre (research grant 1636), Mater Health Services Brisbane Ltd., Australia, and by an unrestricted donation from Laerdal Inc.

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

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](http://www.simulationinhealthcare.com)).

This study was completed in the School of Information Technology and Electrical Engineering at The University of Queensland.

Copyright © 2011 Society for Simulation in Healthcare  
DOI: 10.1097/SIH.0b013e31821eaa38

assessments. Previous healthcare studies<sup>10–12</sup> have shown that judgment analysis techniques can model judgments in different clinical settings. Together, Brunswik's theory and judgment analysis can provide important insight into clinical decision making.<sup>13</sup> An example from an intensive care unit setting illustrates this.

By observing a patient directly and using monitoring equipment, an intensive care unit nurse observes a variety of cues that reflect the clinical state of a patient. The extent to which each cue represents the patient's illness is not explicitly defined, and there may be redundancies in the information presented by the cues. Being attuned to the probabilistic nature of the environment, the nurse uses knowledge and skill to integrate the cues and to make judgments, together comprising an overall assessment of the patient's condition. The nurse then decides how to act to maintain or improve the clinical state of the patient. Different nurses may make different judgments, but training and experience are likely to reduce the variation, and improve accuracy, which in turn has a direct influence on the quality of care provided.

Brunswik's probabilistic functionalism provides a framework to study how an individual, such as the nurse, achieves a consistency of judgment policy that carries over from patient to patient. Judgment analysis provides techniques to quantify each person's judgment policy, allowing their assessment to be predicted in any given situation and allowing changes as a result of training to be measured. For more details about Brunswik's theory and judgment analysis, see Hammond and Stewart,<sup>14</sup> Cooksey,<sup>15</sup> and Kirlik.<sup>13</sup>

To illustrate the relationship between accuracy of clinical assessments and effectiveness of teamwork, we move from the bedside nurse example to complex emergency situations, where high-quality care can be provided only by a well-coordinated team of clinicians.<sup>2</sup> In these situations, each clinician's capacity to monitor cues diminishes because of the focus needed for specific tasks in which he or she is involved. To compensate, team members should share information, so that all members are aware of the details they need to make good decisions and provide appropriate care. Therefore, in emergency situations, the accuracy of clinical assessments that each team member makes depends not only on his or her individual knowledge and skill, as in the example of the bedside nurse, but also on how effectively relevant clinical cues are communicated between all members of the team. Thus, each member's accuracy of assessment depends on the quality of teamwork, meaning that accuracy may be a good indicator for the effectiveness of team coordination.

## ACCURACY SCORE

We measured the accuracy of clinical assessments with a version of the Accuracy Score (AS). The AS<sup>15</sup> is an indexed measure of how accurate a set of judgments is with respect to a set of actual environmental situations. The AS is computed from the discrepancy, or squared error, between the values of the assessments and the actual situations ( $MSE_Y$ ), as a proportion of the maximum squared error possible ( $MSE_{max}$ ), as presented in Eq. (1).

$$AS = 1 - \frac{MSE_Y}{MSE_{max}} \quad (1)$$

The value of the AS can range between 0 (poorest possible accuracy) and 1 (perfect accuracy). More details about the computation of the AS and the derivation of  $MSE_Y$  and  $MSE_{max}$  are presented in Appendix A.

For this article, the AS was used to measure the accuracy of a set of clinical assessments, made during neonatal resuscitation, with respect to the objective illness severity that a mannequin was programmed to represent. The clinical assessments and the objective illness were indicated by Apgar scores.<sup>16</sup> The objective illness severity was computed algorithmically from the values of five clinical signs exhibited by the mannequin. These signs and their theoretical relationship to the computation of Apgar scores for real babies are presented in Appendix B.

We propose that the accuracy of clinical assessments made during simulated clinical emergency scenarios can be computed and can then serve as a measure for the effectiveness of the interaction between team members. Such a measure might then be used to evaluate how factors such as different levels of task demand and training method influence teamwork effectiveness.

## FROM THEORY TO APPLICATION

### Area of Practice

We focused our initial research on neonatal resuscitation. There are three characteristics of neonatal resuscitation that made it suitable as a first setting to test the proposed measure.

First, neonatal resuscitation is an intense emergency activity during which observations and clinical procedures need to be performed in timeframes of minutes and even seconds.<sup>17</sup> As a result, teamwork is expected to be a dominant factor in outcomes.

Second, at the time of this research, the Neonatal Resuscitation Program (NRP) that is taught in many countries worldwide<sup>18</sup> did not include a specific curriculum for team training.<sup>19</sup> We expected that the clinicians who participated in the experiment, and who were previously trained in a course based on NRP, might show noticeable changes in their performance in response to training that was specifically designed to promote teamwork.<sup>19–21</sup>

Third, the Apgar score is a universal system, familiar to all clinicians who care for newborns, for describing the physiological status of a newborn. It is directly computed from numeric values attributed to five types of clinical signs. Although the reliability of the Apgar score has been questioned over the years,<sup>22,23</sup> in previous research, we found that individual clinicians could reliably and accurately<sup>24</sup> interpret clinical signs of the SimNewB (Laerdal Inc.)<sup>25</sup> mannequin from video-recorded simulated resuscitations.

### Hypotheses

Previous studies<sup>21,26,27</sup> have found that team performance is affected by variations in workload and in training. We hypothesized that the AS would be sensitive to similar variables.

The first hypothesis was that when clinicians are engaged in the hands-on task of simulated resuscitation, their Apgar

assessments would be less accurate than when they watch videos of resuscitations performed by others and then make assessments.<sup>26</sup>

The second hypothesis was that Apgar assessments by clinicians trained in Crisis Resource Management (CRM) principles<sup>28</sup> would be more accurate than those by clinicians exposed only to further clinical resuscitation training<sup>21</sup> (later referred as the “Clin-Sim” condition).

The third hypothesis was that Apgar assessments by clinicians exposed to simulator-based CRM training (the “CRM-Sim” condition) would be more accurate than those by clinicians exposed to lecture-based CRM training<sup>27</sup> (the “CRM-Lec” condition).

## METHOD

### Participants

The participants were 17 clinicians, comprising eight neonatal nurses and nine doctors. The doctors included two neonatologists, three neonatal fellows, and four pediatric registrars (approximately equivalent to postgraduate year 2–3 residents) from Mater Mothers’ Hospital (MMH) in Brisbane, Australia. All the participants worked together at MMH and performed neonatal resuscitations as part of their routine scope of practice.

Participants were divided into three experimental groups, and each group was exposed to a different training condition. In each training condition, there was at least one experienced doctor (a neonatologist or neonatal fellow) and one nurse with several years experience in neonatal nursing:

- In the Clin-Sim condition (N = 5), there were three doctors and two nurses (originally the Clin-Sim condi-

tion had three nurses, but one nurse had to withdraw from the experiment before it started).

- In the CRM-Lec condition (N = 6), there were three doctors and three nurses.
- In the CRM-Sim condition (N = 6), there were three doctors and three nurses.

Four of the clinicians indicated that they had prior experience with simulator-based training. One of these was in the Clin-Sim condition, two in the CRM-Lec condition, and one was in the CRM-Sim condition.

The study was approved by the Human Research Ethics Committees of Mater Health Services (Brisbane), The University of Queensland, and Queensland Health.

### Apparatus

The experiment took place in the Queensland Health Skills Development Centre in Brisbane, Australia. The experimental environment included a resuscitation area resembling the MMH neonatal resuscitation environment and a control room separated by one-way viewing glass. The vital signs from a Laerdal SimNewB mannequin were manipulated from the control room and were recorded into dedicated data files. An audiovisual recording system was used to record the scenarios.

## EXPERIMENTAL PROCEDURES

Participants in each training condition spent 2 nonconsecutive days in the Queensland Health Skills Development Centre and progressed in their training condition group through the five phases shown later (further detail in Table 1). The experimental procedure was identical for all participants except during

**Table 1.** Steps in the Experimental Procedure

Experimental Step*	Individual Participant Activities	Differences and Similarities Across Training Conditions
1. Familiarization	1a. Is introduced to the simulated Neonatal Resuscitation (NNR) environment and to the mannequin’s capabilities. 1b. Practices clinical procedures using the mannequin (ventilation, intubation, chest compression, etc.).	Same across conditions. All participants within a condition performed this step together.
2. Baseline assessment	2a. Views 40 2-minute video recordings of simulated NNR scenarios. 2b. After viewing each of the 40 video-recorded NNR scenarios assigns an Apgar score for the mannequin.	Same across conditions. All participants within a condition performed this step at the same time but made their ratings independently.
3. Pretraining assessment	3a. Participates as a team member in nine hands-on 5-minute NNR scenarios. 3b. Individually assigns an Apgar score to the mannequin after each of the nine scenarios.	Pretraining done in three-person teams formed from different combinations of the five or six participants in each condition. Each team had at least one doctor and at least one nurse. Team composition changed after each scenario; 18 different compositions if n = 6; nine different compositions if n = 5.
4. Training	4a. Undergoes training specific to a condition: Clin-Sim: Clinical content and simulator-based hands-on scenarios CRM-Lec: Teamwork content and lecture-based delivery format CRM-Sim: Teamwork content and delivery included simulator-based hands-on scenarios.	Different across training conditions. Further details about training that participants in each condition underwent are in Table 2.
5. Posttraining assessment	5a. Participates as a team member in nine hands-on 5-minute NNR scenarios. 3b. Individually assigns an Apgar score to the mannequin after each of the nine scenarios.	Posttraining done in three-person teams formed from different combinations of the five or six participants in each training condition. Each team had at least one doctor and at least one nurse. Team composition changed after each scenario; 18 different compositions if n = 6; nine different compositions if n = 5.

Steps are shown as the flow of activities, which was identical for each participant (left column). Differences in the procedure across training conditions are shown in the right column.

\*Steps 1–3 and about half of step 4 activities were conducted on the first experimental day; the rest of the activities were conducted on the second day.

the training phase. Training is described in more detail in a later section.

1. Familiarization: The simulation environment and procedures were fully explained to participants.
2. Baseline: The participants viewed video-recorded scenarios and provided Apgar assessments (additional details about the preparation and presentation of the recordings were reported in Nadler et al<sup>24</sup>). This phase was required to make sure that the three experimental groups were balanced for factors such as individuals' clinical experience and familiarity with simulated tasks.
3. Pretraining (provide Apgar scores after hands-on scenarios): Each participant took part in nine hands-on resuscitation scenarios in rotating teams of three and provided Apgar assessments at the end of the scenario. There was always at least one doctor and one nurse in each team. Team structure is described further in Table 1.
4. Training (Clin-Sim, CRM-Lec, or CRM-Sim): Training is described later and in Table 2.
5. Posttraining (provide Apgar scores after hands-on scenarios): The same procedure was used as in the pretraining phase.

Before the start of each hands-on scenario in the pretraining and posttraining phases, the participants were given a brief clinical history for the simulated patient. Once a scenario started, the simulator was manipulated from the control room to exhibit naturalistic responses of the mannequin (either improvement or deterioration) in response to the clinicians' activities. Regardless of the clinical status of the mannequin, hands-on scenarios were stopped after 5 minutes. A recording from one of these scenarios is provided as Video, Supplemental Digital Content 1, <http://links.lww.com/SIH/A21>.

### Scenario Design

The video-recorded scenarios that were presented in the baseline phase were designed to present a large variety of situations, showing the mannequin in different illness states. The design was guided by Brunswik's theory and by the needs of the judgment analysis framework. The design and the validation of these recordings were reported in detail in the study by Nadler et al.<sup>24</sup>

The scenarios in the pretraining and posttraining phases each started with a predefined set of clinical signs (eg, for heart rate, breathing, oxyhemoglobin saturation, vocal sounds, and movement). Each scenario started with different clinical signs, but the values of the clinical signs at the start of the scenarios were matched in the three groups.

### Training Conditions

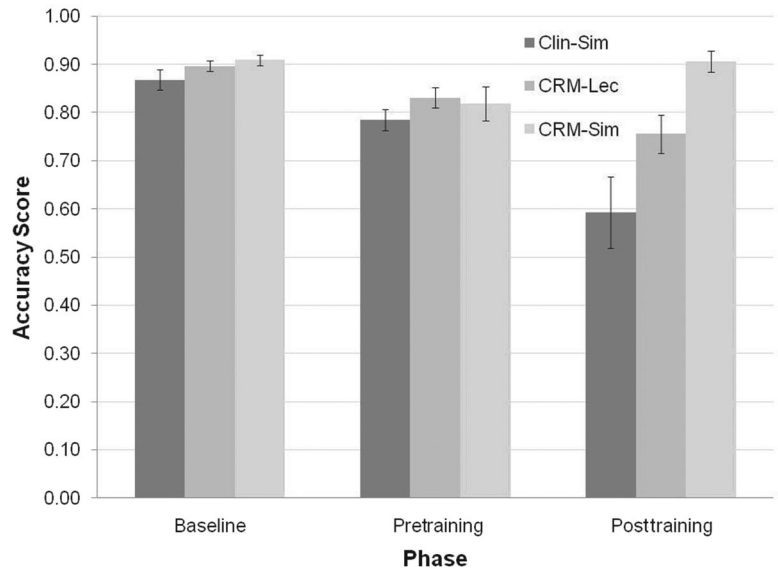
During the training phase, participants in each training condition experienced a different training intervention, as follows. Further details are provided in Table 2.

1. Participants in the Clin-Sim condition underwent clinical training, which included lectures, hands-on resuscitation scenarios, and debriefings. There was no focus on working in teams.
2. Participants in the CRM-Lec condition underwent training that included lectures about CRM and video presentations of CRM. This training did not include hands-on resuscitation scenarios.
3. Participants in the CRM-Sim condition underwent training that included presentations about CRM, video demonstrations of CRM, hands-on resuscitation scenarios to allow participants to practice applying CRM principles, and debriefings in which teamwork was discussed.

**Table 2.** Detailed Summary of the Interventions that Participants in Each Training Condition Experienced in the Training Phase of the Experiment

	Clin-Sim	CRM-Lec	CRM-Sim
Training goal	Reinforcement of neonatal resuscitation knowledge and clinical practice	Introduction to team-related skills in neonatal resuscitation context	Introduction to team-related skills in neonatal resuscitation context
Topics emphasized	Newborn physiology, clinical complications, resuscitation equipment, clinical procedures, application of drugs	Leadership, communication, allocating attention, workload distribution, anticipating and planning, early call for help	Leadership, communication, allocating attention, workload distribution, anticipating and planning, early call for help
Delivery method	Lectures Hands-on scenarios	Lectures Video demonstrations—presentation and evaluation	Lectures Video demonstrations Hands-on scenarios
Sources for clinical training materials	Presentations based on the Mater Mothers' Hospital neonatal resuscitation training	Summary of Mater Mothers' Hospital neonatal resuscitation training	Summary of Mater Mothers' Hospital neonatal resuscitation training
Sources for team training materials	N/A	Modified presentation from a maternity Crisis Resource Management course	Modified presentation from a maternity Crisis Resource Management course
Instructors	Educational nurse from the Mater Mothers' Hospital	Educational nurse (same as in Clin-Sim condition), maternity CRM course instructor from Queensland Health Skills Development Centre	Educational nurse (same as in Clin-Sim condition), maternity CRM course instructor from Queensland Health Skills Development Centre
Number of (time spent on) hands-on training scenarios and debriefings	4 (2:10 h)	0	3 (2:00 h)
Overall duration	Total of 6:00 training hours	Total of 6:00 training hours	Total of 6:00 training hours

**Figure 1.** Accuracy Scores (means and standard errors of the mean) for the three training conditions in the three experimental phases. An Accuracy Score of 1.0 indicates perfectly accurate Apgar score judgments whereas 0.0 indicates poorest possible accuracy. [Note: Baseline data were reported in Nadler I, Liley HG, Sanderson PM. Clinicians can accurately assign Apgar scores to video recordings of simulated neonatal resuscitations. *Simulation in Healthcare* 2010;5:204–212.]



## Analysis

### Accuracy Score

ASs were computed for each participant for the baseline, pretraining and posttraining phases of the experiment. For the computation, we compared the Apgar score judgments that each clinician made after each scenario and the known, objective state of the mannequin (in Apgar score terms) at the end of each scenario. As explained earlier, the objective state was computed from five clinical signs that the mannequin presented. The data sets that were used to calculate  $MSE_Y$  and  $MSE_{max}$  are presented in Appendix A, Table A1.

### Statistics

For the first analysis step, we used a two-way analysis of variance with the between-subjects factor of training conditions (three levels: Clin-Sim, CRM-Lec, and CRM-Sim) and the within-subjects factor of phases (three levels: baseline, pretraining, and posttraining). At the second analysis step, we used a post hoc Tukey Honestly Significant Difference (HSD) test to evaluate differences between the means for each of the three phases and a post hoc Tukey unequal-n HSD test to evaluate differences between all nine cells in the design. All statistical testing was performed with Statistica (version 9.0). Results for the baseline phase were also reported in a previous publication, where they were analyzed for other purposes.<sup>24</sup>

## RESULTS

We found significant differences between training conditions ( $P < 0.01$ ), between phases ( $P < 0.001$ ) and between the interaction of training conditions and phases ( $P < 0.001$ ) (Fig. 1).

The ASs decreased significantly from the baseline phase to the pretraining phase ( $P < 0.01$ , post hoc Tukey HSD). Between these phases, task demand increased, impairing clinicians' capability to monitor closely the mannequin's clinical state and resulting in less accurate assessments. This result was consistent with our first hypothesis.

In the posttraining phase, the ASs for the CRM-Sim condition and the CRM-Lec condition were each significantly higher than for the Clin-Sim condition ( $P < 0.001$  and  $P <$

0.05, respectively, all comparisons by post hoc Tukey unequal-n HSD test), consistent with our second hypothesis that CRM training would enable better assessments than clinical training. However, unexpectedly, the AS for the Clin-Sim condition decreased significantly from the pretraining to the posttraining phase ( $P < 0.01$ ).

During the posttraining phase, the AS for the CRM-Sim condition was significantly higher than for the CRM-Lec condition ( $P < 0.05$ , post hoc Tukey unequal-n HSD test). This was consistent with our third hypothesis, which was that there would be an advantage of CRM training with hands-on practice over CRM training with no hands-on practice.

## DISCUSSION

The results show that the AS was sensitive to all the variations of the experiment. The differences in the AS between the three training conditions were all in the direction expected, in the light of previous studies on team training.<sup>21,26,27</sup>

The significant decrease in the AS from the baseline phase, where participants provided Apgar scores for video-recorded scenarios performed by others, to the pretraining phase, where participants provided Apgar scores for scenarios in which they took part, showed sensitivity to task demand. The similar decrease in the AS for all three training conditions between these two phases suggests that the groups were balanced for variables such as clinical experience that otherwise could have biased the results.

The AS showed that clinicians who underwent any CRM training were more accurate in their posttraining clinical assessments than clinicians who underwent only clinical NRP-based training. However, this finding was influenced more by the significant decrease in the AS for the Clin-Sim condition from the pretraining to posttraining phase than by improvement in the CRM training conditions, even though results for the CRM-Sim condition showed a slight trend to improve after training. All five Clin-Sim participants provided less accurate Apgar assessments after training, and for two doctors and one nurse the effect was quite marked. A possible explanation for the decrease could be self-imposed pressure

to execute the clinical material just learned, which could have made participants focus on adjusting their knowledge and skills at the expense of the quality of their teamwork. Regardless of cause, this decrease could be transient, and we cannot conclude that there will be a long-term advantage of CRM training over clinical training for accuracy of clinical assessments. Clearly, further research is needed to understand the decrease in AS for the Clin-Sim condition after training and to test the generality of the results for all three training conditions.

The superior accuracy after the CRM-Sim training when compared with the CRM-Lec training indicates that the AS was sensitive to differences in CRM training with versus without hands-on practice and supports expectations that hands-on practice is important in improving teamwork. The accuracy for the CRM-Sim condition was as high in the post-training phase as in the baseline phase, in contrast to the other two training conditions. This might indicate that the CRM-Sim team training helped clinicians to compensate for the higher task demand of hands-on resuscitations compared with simply observing recorded scenarios. The superiority of the CRM-Sim condition could be due to better team communication, where updated information is being shared more effectively between team members, so they can make their judgments with more up-to-date information. Alternatively, role allocation, which was also emphasized in the CRM training, could have improved, enabling one or more team members to make more frequent and accurate observations and convey the information to the rest of the team.

Nevertheless, although the AS changed with changes in task demand and training condition, in the absence of well-established models to explain how team members share different attributes of a situation,<sup>29</sup> we do not know exactly why the AS changed. Rather than measuring team performance by assessing qualitative or quantitative aspects of team members' interactions, the AS reflects how accurately clinicians assess a patient's clinical state, which is critical for decision making about interventions and, ultimately, for patient outcomes.<sup>12</sup> The AS is, therefore, inherently different from other measures of team performance, and it is particularly meaningful to healthcare because it reflects diagnostic accuracy.

There are advantages when computing the AS based on Apgar scores. The Apgar calculation is simple and universal, and the component signs for simulated patients resemble those used for decision making during actual neonatal resuscitation. These characteristics made the Apgar score an excellent candidate for applying Brunswik's probabilistic functionalism and judgment analysis.

The results of this study apply not only to neonatal resuscitation but also potentially to other clinical emergency areas for which simulator-based training is used. Scenarios would need to be prepared and validated using the method described in Nadler et al.<sup>24</sup> Application to other clinical areas would require a well-established clinical score that can be used for computing the actual state of the simulated patient.

### Limitations and Future Studies

This study presents an initial test of the AS as a measure of the effectiveness of simulator-based team training. There are

several limitations of the study that need to be addressed in future research to determine the validity and generality of the results.

First, the reliability and validity of the AS should be tested in studies with a larger number of participants. Our small sample was due mainly to the small number of participants available to us at the time.

Second, the AS should be tested in a study that uses multiple independent teams with fixed compositions within and across the pretraining and posttraining phases, rather than changing compositions as in this study. With fixed-composition teams, the effects of team training would be tested on independent rather than partially dependent samples. Moreover, fixed-composition teams might develop their own strategies<sup>26</sup> when applying newly learned material, and the impact of team training on the AS could be more evident as a result.

Third, from the present investigation, we cannot be completely certain that the AS reflects team effectiveness. We did not have a control condition in which no training was provided between the pretraining and posttraining assessments, which could have helped us to interpret the unexpected Clin-Sim result. Some changes in the AS could have been due to external factors and not to the training manipulations. Furthermore, we cannot determine from our experiment the exact mechanisms through which hands-on team training led to better clinical assessments. In future studies, results based on the computation of the AS should be compared with results based on other validated measures of team training effectiveness.

Fourth, the calculation of the AS requires that the illness severity of the patient can be represented with a well-established clinical score (or criterion in Brunswik's terms). This research made use of the Apgar score, but Apgar scores apply only to newborn infants, and in their simplicity, relevance and familiarity could be uniquely suited to our approach. To date, judgment analysis studies in other clinical areas have quantified clinicians' assessments by establishing criteria either from predictive models<sup>30</sup> or from published "gold standard" computations.<sup>11</sup> However, the usefulness of the AS as an indicator of teamwork effectiveness needs to be tested in other clinical realms and with measures other than the Apgar score.

## CONCLUSIONS

Currently, there is no standard measure or method for evaluating team training effectiveness.<sup>4</sup> The results of this initial study suggest that the AS may be a useful additional measure for testing the effectiveness of teamwork training in simulated scenarios. Our research has opened many questions and to date has left many unanswered, but if future studies indicate that the AS is reliably sensitive to team training manipulations, it may prove to have several advantages over measures of team effectiveness that require scoring by observers. The computation of the AS does not pose a burden on any observers who might be present and does not require off-line viewing and scoring.<sup>2</sup> The AS could increase the objectivity of team training evaluation<sup>31</sup> and consequently in-

crease its reliability and validity. When used alongside other measures,<sup>2,4</sup> the AS could make the evaluation of training interventions more comprehensive and thorough.

Brunswik's probabilistic functionalism provides a solid theoretical basis for explaining how assessments are made during the performance of complex clinical interventions and judgment analysis leads us to the specific formula for the AS. Changes in the AS may serve as a measure for evaluating the effectiveness of team training interventions. Overall, the approach may provide a new framework for studies about training in healthcare<sup>13</sup> that ultimately improve patient outcomes.

## APPENDIX A: METHOD FOR CALCULATING THE AS

The Accuracy Score (AS) is presented in Eq. (2)

$$AS = 1 - \frac{MSE_Y}{MSE_{max}} \quad (2)$$

$MSE_Y$  = mean of the squared errors between a set of assessments and the reference set.

$MSE_{max}$  = maximal squared error between a set of assessments and the reference set.

We outline the specific methods used to calculate  $MSE_Y$  and  $MSE_{max}$ , especially because there are alternatives for calculating the latter in particular.

$MSE_Y$  can be calculated by using Eq. (3)<sup>32</sup>:

$$MSE_Y = (\bar{J} - \bar{O})^2 + (s_j - s_o)^2 + 2s_j s_o (1 - r_{JO}) \quad (3)$$

where:

$\bar{J}$ , mean of the judged scores

$\bar{O}$ , mean of the objective scores

$s_j$ , standard deviations of the judged scores

$s_o$ , standard deviations of the objective scores

$r_{JO}$ , correlation between the judged and the objective scores

There are different methods for estimating  $MSE_{max}$ . One option is to use the squared value obtained by subtracting the most extreme values in the scores' scales. Cooksey<sup>15</sup> indicates that such a calculation is too conservative, and a more realistic approach is to calculate  $MSE_{max}$  values obtained from the two sets of scores. This calculation is presented in Eq. (4).

$$MSE_{max} = \{[\min(J_{max}, \bar{J} + 2s_j)] - [\max(O_{min}, \bar{O} - 2s_o)]\}^2 \quad (4)$$

where:

$J_{max}$ , highest score out of all the judged scores

$O_{min}$ , lowest score out of all the objective scores

In experiments such as ours, in which participants were exposed to the entire possible range of situations, this approach and the previous one become practically identical. In both cases,  $MSE_{max}$  includes the highest potential error values, so it is large with respect to  $MSE_Y$ . This limits the possible values for the AS to the high end of the scale. To overcome these limitations, we calculated  $MSE_{max}$  in a similar manner to that suggested by Cooksey, but instead of using judged scores and objective scores, we used the actual error values. Equation 5 represents the  $MSE_{max}$  that we used.

An error in observation  $i$  is defined as:

$$e_i = J_i - O_i$$

$$MSE_{max} = [\min(|e|_{max}, |\bar{e}| + 2s_e)]^2 \quad (5)$$

Where:

$|e|_{max}$ , highest absolute value of all errors

$|\bar{e}|$ , mean absolute value of all errors

$s_e$ , standard deviations of the values of the errors.

Table A1 lists data sets used for the computation of the AS for each clinician in each of the experimental phases.

**Table A1.** Data Sets that were Used to Calculate the Accuracy Scores for the Three Phases

Phase	Data Sets for Calculating Each Clinician's Mean Square Error ( $MSE_Y$ )	Data Sets for Calculating the Maximal Square Error ( $MSE_{max}$ ) for All Clinicians in Each Training Condition
Baseline	30* Apgar score judgments for the scenarios presented. 30* objective Apgar scores calculated for the scenarios presented.	510 errors between the judged values and the objective values (17 clinicians, 30 error values per clinician)
Pretraining and posttraining	9 Apgar score judgments for the hands-on scenarios in each of the two phases. 9 objective Apgar scores calculated for the hands-on scenarios in each of the two phases.	For Clin-Sim condition, 45 errors (5 clinicians and 9 errors per clinician) For CRM-Lec and CRM-Sim conditions, 54 errors (six clinicians and 9 errors per clinician)

\*Ten of the 40 recordings presented in the baseline phase showed the mannequin with rapid respiration rates that can be typical for sick infants but not immediately after birth. The rest of the scenarios in the experiment showed the mannequin with clinical signs that are typical immediately after birth. This difference could potentially lead to incorrect Apgar score assessments, and therefore, the scores of these 10 recordings were excluded from the computation of the Accuracy Score.

## APPENDIX B: COMPUTATION OF AN APGAR SCORE FOR THE SimNewB™ AND FOR A REAL BABY

The Apgar score is computed from five clinical signs that contribute 0, 1, or 2 points to the total score that can range

from 0 to 10. Table B1 presents the contribution of each clinical sign according to its state when used for computing the Apgar scores for real babies and the equivalent signs when used for the mannequin.

**Table B1.** Comparison Between the Contributions of the Clinical Signs When Used for Assigning Apgar Scores to Babies 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 breaths/min	>30 breaths/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%

Adapted with permission from *Simul Healthc.* 2010;5:204–212.

### ACKNOWLEDGMENTS

This article was written as part of Izhak Nadler's PhD studies at the University of Queensland while he held Endeavor IPRS and UQILAS scholarships. The authors acknowledge Parker Healthcare's loan of the ATOM cot and of Midmed's loan of the resuscitation trolley. The authors are grateful for the participation and in kind support from management and staff of the Queensland Health Skills Development Centre and from the Mater Mothers' Hospital. The authors thank all the clinicians from the Mater Mothers' Hospital who volunteered to participate in the experiment. The authors acknowledge the advice and support of Ray Cooksey with respect to judgment analysis, but any errors are the authors' responsibility. Finally, they thank Coleen van Dyken, Daniel Host, and Pauline Lyon for their contributions to the preparation and to the conduct of this study.

### REFERENCES

- Burke CS, Salas E, Wilson-Donnelly K, Priest H. How to turn a team of experts into an expert medical team: guidance from the aviation and military communities. *Qual Saf Health Care* 2004;13:196–1104.
- Rosen MA, Salas E, Wilson KA, et al. Measuring team performance in simulation-based training: adopting best practices for healthcare. *Simul Healthc* 2008;3:33.
- Salas E, Rosen MA, Weaver SJ, Held JD, Weissmuller JJ. Guidelines for performance measurement in simulation-based training. *Ergon Des Q Hum Factors Appl* 2009;17:12–18.
- Salas E, Rosen MA, Held JD, Weissmuller JJ. Performance measurement in simulation-based training: a review and best practices. *Simul Gaming* 2009;40:328.
- Flin R, Maran N. Identifying and training non-technical skills for teams in acute medicine. *Qual Saf Health Care* 2004;13:i80–i84.
- Fletcher G, Flin R, McGeorge P, Glavin R, Maran N, Patey R. Anaesthetists' Non-Technical Skills (ANTS): evaluation of a behavioural marker system. *Br J Anaesth* 2003;90:580–588.
- Flin R, Fletcher G, McGeorge P, Glavin R, Maran N, Patey R. Rating anaesthetists' non-technical skills: the ANTS system. *Proceedings of the 47th Human Factors and Ergonomics Society Conference*; 2003:1498–1506.
- Brunswik E. *Perception and the Representative Design of Psychological Experiments*. 2nd ed. Berkeley, CA: University of California; 1956.
- Brunswik E. *The Conceptual Framework of Psychology*. Chicago, IL: University of Chicago Press; 1952.
- Jacklin R, Sevdalis N, Darzi A, Vincent C. Mapping surgical practice decision making: an interview study to evaluate decisions in surgical care. *Am J Surg* 2008;195:689–696.
- Jacklin R, Sevdalis N, Harries C, Darzi A, Vincent C. Judgment analysis: a method for quantitative evaluation of trainee surgeons' judgments of surgical risk. *Am J Surg* 2008;195:183–188.
- Thompson C, Bucknall T, Estabrookes CA, et al. Nurses' critical event risk assessments: a judgement analysis. *J Clin Nurs* 2009;18:601–612.
- Kirlik A. Brunswikian theory and method as a foundation for simulation-based research on clinical judgment. *Simul Healthc* 2010;5: 255–259.
- Hammond KR, Stewart TR. *The Essential Brunswik: Beginnings, Explications, Applications*. New York: Oxford University Press; 2001.
- Cooksey RW. *Judgment Analysis: Theory, Methods and Applications*. San Diego: Academic Press; 1996.
- Apgar V. A proposal for a new method of evaluation of the newborn infant. *Curr Res Anesth Analg* 1953;32:260–267.
- Kattwinkel J, Short J. *Textbook of Neonatal Resuscitation*. Elk Grove Village, IL: 5th ed. American Academy of Pediatrics; 2006.
- AAP. *Neonatal Resuscitation Program*. Available at: [http://www.aap.org/nrp/intl/intl\\_where.html](http://www.aap.org/nrp/intl/intl_where.html). Accessed December 23, 2010.
- Thomas EJ, Sexton JB, Lasky RE, Helmreich RL, Crandell DS, Tyson J. Teamwork and quality during neonatal care in the delivery room. *J Perinatol* 2006;26:163–169.
- Halamek LP, Kaegi DM, Gaba DM, et al. Time for a new paradigm in pediatric medical education: teaching neonatal resuscitation in a simulated delivery room environment. *Pediatrics* 2000; 106:E45.
- Thomas EJ, Taggart B, Crandell S, et al. Teaching teamwork during

- the Neonatal Resuscitation Program: a randomized trial. *J Perinatol* 2007;27:409–414.
22. Papile LA. The Apgar score in the 21st century. *N Engl J Med* 2001;344:519.
  23. Pinheiro JM. The Apgar cycle: a new view of a familiar scoring system. *Arch Dis Child* 2009;94:F70–F72.
  24. Nadler I, Liley HG, Sanderson PM. Clinicians can accurately assign Apgar scores to video recordings of simulated neonatal resuscitations. *Simul Healthc* 2010;5:204–212.
  25. Laerdal. SimNewB™. <http://www.laerdal.com/document.asp?subnodeid=32779467>. Accessed December 23, 2010.
  26. Adelman L, Miller SL, Henderson DV, Schoelles M. Using Brunswikian theory and a longitudinal design to study how hierarchical teams adapt to increasing levels of time pressure. *Acta Psychol* 2003;112:181–206.
  27. Smith-Jentsch KA, Salas E, Baker DP. Training team performance-related assertiveness. *Pers Psychol* 1996;49:909–936.
  28. Gaba DM, Fish KJ, Howard SK. *Crisis Management in Anesthesiology*. New York: Churchill Livingstone; 1994.
  29. Salas E, Cooke NJ, Gorman JC. The science of team performance: progress and the need for more. *Hum Factors* 2010;52:344–346.
  30. Beckstead JW, Stamp KD. Understanding how nurse practitioners estimate patients' risk for coronary heart disease: a judgment analysis. *J Adv Nurs* 2007;60:436–446.
  31. Salas E, Cooke NJ, Rosen MA. On teams, teamwork, and team performance: discoveries and developments. *Hum Factors* 2008;50:540–547.
  32. Skinner HA. Differentiating the contribution of elevation, scatter and shape in profile similarity. *Educ Psychol Meas* 1978;38:297.