# Using Brunswik's Probabilistic Functionalism to Test How Clinicians Make Judgments in Simulated Neonatal Resuscitation Scenarios

Izhak Nadler[1], Penelope M. Sanderson[1,2]
[1]School of Information Technology and Electrical Engineering
[2]School of Psychology and School of Medicine
The University of Queensland, Brisbane, Australia

Accurate clinical assessments are essential for providing appropriate patient care. However, clinicians do not necessarily make accurate assessments, either because they are overloaded with other tasks, or because the computation of assessments is ambiguous. We used Brunswik's *probabilistic functionalism* to study how clinicians assessed the clinical state of a mannequin when they viewed recordings of simulated neonatal resuscitation scenarios and when they performed simulated resuscitations. Seventeen clinicians individually assigned an Apgar (neonate illness) score to 30 pre-recorded scenarios and also to 9 scenarios in which they played a hands-on role. We computed a *judgment policy* for each clinician showing the relative importance of five clinical signs that constitute the Apgar score. The accuracy of clinicians' judgment policies was significantly correlated with the accuracy of their Apgar assessments for the pre-recorded scenarios ($p<0.01$) but not for the hands-on scenarios. The weighting for the clinical signs in the judgment policies was different from the unit weighting in the Apgar score itself. Brunswik's approach provided a useful framework for testing clinicians' assessments in simulated neonatal resuscitations. Future studies should determine the factors that affect accuracy in hands-on scenarios and test the applicability of the methods presented for other healthcare practice areas.



## INTRODUCTION

Accurate clinical assessments are necessary for initiating correct clinical treatment, especially in the case of medical emergencies (Greenland et al., 2007; Thompson et al., 2009). Nonetheless, Greenland et al., Thompson et al., and other studies (Beckstead & Stamp, 2007; Wigton, 2008) show that clinical information is not used systematically when clinicians make their assessments. Two possible reasons are (1) clinicians may be overloaded with other tasks and therefore may make general assessments rather than accurate ones and (2) data that are used for the assessment may be subjective or not very clearly defined. To increase the accuracy of clinical assessments, we need to study how clinicians make assessments and identify factors that may limit the accuracy of their assessments.

Judgment Analysis (JA) is a research area in which the accuracy of judgments is modeled, quantified and analyzed. Brunswik's *probabilistic functionalism* (Brunswik, 1956) has provided the theoretical framework for many JA studies that have tested individual judgments in many domains (Hammond & Stewart, 2001). In a recent review, Kirlik (2010) indicated that Brunswik's approach has proven to be useful in studies in aviation and other safety critical industries. Kirlik noted that researchers in healthcare may find this approach useful as well.

Brunswik's Lens Model (Figure 1) illustrates how a person makes a judgment ($Y_s$) of an environmental criterion or situation ($Y_e$) by observing and weighting ($b_i$) cues ($X_i$) that represent the situation with different validities ($w_i$).

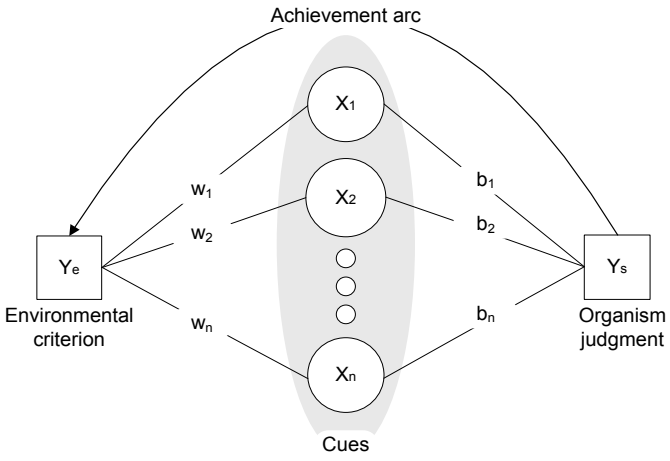


Figure 1: Brunswik's Lens Model showing the environmental criterion ($Y_e$), ecological validities of cues ($w_i$), cue values ($X_i$), relative weights that the person unconsciously associates with each cue ($b_i$), and the person's judgment ($Y_s$)

According to Brunswik, an individual observes the cues that emerge from the environment and then uses skills and knowledge to interpret the cues and make a judgment. After judging the situation the person can act and respond as required.

In an earlier analysis of part of the dataset presented here, we tested the accuracy of clinicians' judgments when they observed recordings of simulated neonatal resuscitations (Nadler, Liley, & Sanderson, 2010). We used Brunswik's *representative design* and JA concepts to design the study. The main conclusions were that clinicians could accurately and systematically assess the clinical state of a mannequin when viewing recordings of simulated neonatal resuscitations. However, it was unknown how they would perform during hands-on resuscitations.

The clinicians used the Apgar score (Apgar, 1953) of neonate illness to indicate the illness severity of the simulated neonate (a mannequin). The computation of the Apgar score for the simulated neonate was based on five clinical signs, similar to the signs that are used to compute an Apgar score for real neonates. Each sign contributes an equal portion to the total score (see Table 1), and according to its state, each sign donates 0, 1 or 2 points to the Apgar score which ranges between 0 (critically ill neonate) and 10 (best possible condition of the neonate).

Table 1: Contribution of five clinical signs to the total Apgar score for a simulated neonate patient

| SIGN | Contribution to the total Apgar score | | |
|---|---|---|---|
| | 0 | 1 | 2 |
| Heart rate | Absent | <100 beats/min | >100 beats/min |
| Respiratory rate | Absent | <30 breath/min | >30 breath/min |
| Oxyhemoglobin saturation | <76% | 76-82% | 83-100% |
| Muscle tone | None | Tone | Movement |
| Vocal sounds | None | Weak cry, hiccup, grunting | Strong, normal cry, or scream |

Pediatric text books clearly indicate that the Apgar score should be computed from the five clinical signs (Baker, 2006; Lowdermilk & Perry, 2006; Murray & McKinney, 2010). At the early stages of this study we reviewed clinical records of actual births at the participating hospital. We noted that the values of the five signs were rarely indicated, possibly suggesting that the score was assigned holistically and not necessarily algorithmically. We are not aware of any study in the literature that has tested such a possibility.

**Goal of Paper**

In this paper we used Brunswik's approach to identify how clinicians assessed the clinical state of a mannequin when they viewed recorded resuscitations and when they performed hands-on resuscitation scenarios. The data were collected as part of a larger study reported in Nadler, Sanderson, & Liley (in press). Our hypotheses for the present study were as follows:

1. The Apgar assessments that neonatal resuscitation clinicians make when viewing recordings of simulated resuscitations can be modeled with a linear equation ('judgment policy').
2. The judgment policies of neonatal resuscitation clinicians may not necessarily resemble the theoretical computation of the Apgar score.
3. Clinicians' ability to assess the clinical state of the mannequin accurately while performing simulated resuscitations will be affected only to a small degree by their ability to monitor and correctly interpret the clinical signs that the mannequin presents.

## METHOD

**Participants**

Participants were nine doctors and eight nurses from the Mater Mother's Hospital in Brisbane, Australia who all perform neonatal resuscitation as part of their routine clinical practice.

**Setup**

The experiment took place in the Queensland Health Skills Development Centre in Brisbane, Australia. The simulated environment was a close replica of the hospital practice environment.

Participants viewed pre-recorded simulated resuscitations before performing the hands-on simulated resuscitations. When the clinicians performed the resuscitations, the simulator was controlled from a control room with one sided viewing glass. All the scenarios performed were recorded and the data were captured in digital log files.

The resuscitations were performed with the SimNewB™ (Laerdal Inc.) mannequin and the clinical signs were presented as follows: (1) heart rate and oxyhemoglobin saturation were presented on a patient monitor, (2) vocal sounds were presented through the mannequin's loudspeaker, (3) chest wall movement reflecting respiration rate and limb position/movement representing muscle tone were monitored by observing the mannequin, (4) cardiac and respiration sounds could be

monitored with a stethoscope, and (5) heart rate could be sensed by palpating the mannequin's umbilical cord.

The recorded scenarios presented the immediate surroundings of the mannequin with an overlay of the patient monitor. The vocal sounds were replayed through a loudspeaker while the cardiac and respiration sounds were played through simulated stethoscopes that each clinician could use.

**Experimental Procedure**

The procedure involved the following steps.

1. Participants were briefed about the experiment and simulator, and they then explored the simulator's capabilities and limitations.
2. Participants viewed and individually assigned an Apgar score to 30 recordings of 2-minute simulated resuscitation scenarios ("viewing" condition).
3. In teams of three members, participants participated in 5-minute resuscitation scenarios. Each participant participated in nine scenarios. Team compositions were changed between the scenarios. When a scenario ended, each participant individually indicated an Apgar score that reflected his/her impression of the clinical state of the mannequin at the end of each scenario ("hands-on" condition).

In a further step, the results of which are reported in Nadler et al. (in press), participants underwent one of three alternative forms of training and then again participated in 5-minute resuscitation scenarios.

The clinicians were not given training in how to assign Apgar scores to the mannequin, nor given a feedback about the accuracy of their judgments.

**Measures and Analysis**

We computed the judgment policy for each participant, which was the relative weight the participant associated with each of the five clinical signs when assessing the Apgar score. We used the JA method (Cooksey, 1996) to capture each judgment that a clinician made with a linear model (see Equation 1):

$$Y_k = \sum_{i=1}^{5} (a_{ik} \times b_i) + e_k \qquad (1)$$

- $Y_k$ – *the value of the judgment in scenario k*
- $a_{ik}$ – *the contribution of sign i (out of the five signs) to the total Apgar score in scenario k*
- $b_i$ – *the relative weight in the policy of sign i*
- $e_k$ – *the error between the actual judgment and the judgment policy in scenario k*

The values of the five $b_i$'s were computed with a regression that minimized the squared errors ($e_k$) over the 30 judgments that a participant indicated in the viewing condition.

An 'Objective Apgar score' was computed for every scenario (viewed or performed) from the values of the five signs that the mannequin presented at the end of the scenario. The Objective Apgar score represents the 'real' clinical state of the mannequin in Apgar score terms.

The 'Judged Apgar score' was the score that each participant indicated at the end of each scenario.

The 'Predicted Apgar score' was computed for each participant for each scenario from that participant's judgment policy and the values of the signs at the end of the scenario, which is the modeled part of the judgment. The formula for computing the Predicted Apgar score for scenario *k* is shown in Equation 2. The Predicted Apgar score was therefore the Apgar score a participant would make if he/she followed his/her own judgment policy perfectly.

$$\sum_{i=1}^{5} (a_{ik} \times b_i) \qquad (2)$$

The Accuracy Score (Cooksey, 1996) is used to show how accurate one set of scores is with respect to another set of scores. It is an indexed measure that ranges between 0 (poorest possible accuracy) and 1 (perfect accuracy). The Accuracy Score (*AS*) is presented in Equation 3.

$$AS = 1 - \frac{MSE_Y}{MSE_{\max}} \qquad (3)$$

- $MSE_Y$ – *the mean of the squared errors between the tested set of scores and the reference set*
- $MSE_{max}$ – *the maximal squared error between the two sets*

We used the *AS* to express how accurate each participant's Judged Apgar scores and Predicted Apgar scores (tested sets) were compared with the (real) Objective Apgar scores (reference set). For more details about the *AS* see Cooksey (1996).

For each participant, we computed the accuracy of their judgments (Judged Apgar scores) with respect to the real state of the mannequin (Objective Apgar scores) to arrive at the 'Judged-Objective Accuracy Score'. We produced a Judged-Objective Accuracy Score for both the viewing condition and the hands-on condition.

Similarly, for each participant, we computed the accuracy of their judgment policy (Predicted Apgar scores) with respect to the real state of the mannequin (Objective Apgar scores) to arrive at the 'Predicted-Objective Accuracy Score'.

To find what portion of the variance in judgments in the viewing condition was explained by participants' judgment policy, we tested the correlation between the

Judged-Objective Accuracy Score and the Predicted-Objective Accuracy Score for the viewed scenarios.

To test what might make some participants' policies more accurate than other participants' policies, we tested the correlation between each of the five normalized weights, $b_i$'s, for the clinical signs in Equation 1 and the Predicted-Objective Accuracy score.

To test if a more accurate policy was associated with more accurate judgments during the hands-on scenario, we tested the correlation between the Judged-Objective Accuracy Score in the hands-on scenario condition and the Predicted-Objective Accuracy Score in the viewing condition. We used a t-test to compare the assessments in the viewing condition and the assessments in the hands-on condition.

## RESULTS

In the viewing condition, the Judged-Objective Accuracy Score and the Predicted-Objective Accuracy Score were significantly correlated ($r$=0.75, $n$=17, $p$<0.01). This means that 56% of the variation in judgments in the viewing condition was explained by the participants' judgment policies. A summary of the *AS* results is presented in Table 2.

Table 2: Summary of Accuracy Scores (*AS*)

| | Mean, (SD), range |
|---|---|
| Judged-Objective *AS* in the viewing condition | 0.89*, (0.04), 0.79-0.93 |
| Judged-Objective *AS* in the hands-on condition | 0.81*, (0.07), 0.69-0.93 |
| Predicted-Objective *AS* by the policies | 0.87, (0.06), 0.77-0.95 |

*Significantly different at p<0.01

The clinical sign that contributed the most to participants' judgment policies, regardless of policy accuracy, was the heart rate. On average, 46% of the assessment that was predicted by the clinicians' judgment policies was based on the heart rate that the mannequin presented. (see Table 3). The other signs had smaller contributions to the judgment policies. Weights for oxyhemoglobin saturation and muscle tone correlated significantly with the accuracy of the policies.

The correlation between the Judged-Objective Accuracy Score in the hands-on condition and the Predicted-Objective Accuracy Score in the viewing condition fell below significance ($r$=0.42, $n$=17, .1>$p$>.05). As reported in Nadler et al. (in press) the accuracy of the assessments in the hands-on condition was significantly lower than in the viewing condition ($p$<0.001)

Table 3: Weights for the clinical signs and their correlation with the accuracy of the judgment policy

| | Average normalized weights of the clinical signs in the judgment policies: Mean, (SD), range | Correlation with the judgment policy |
|---|---|---|
| Heart rate | 0.46, (0.09), 0.32-0.65 | -.30 |
| Respiratory rate | 0.14, (0.06), 0.03-0.26 | -.21 |
| Oxyhemoglobin saturation | 0.13, (0.09), -0.02-0.35 | -.58* |
| Muscle tone | 0.18, (0.10), 0.01-0.36 | .52** |
| Vocal sounds | 0.10, (0.06), 0.01-0.24 | .27 |

*Significant at p<0.01
** Significant at p<0.05

## DISCUSSION

The assessments that the clinicians made while viewing recordings of simulated resuscitation were mainly derived from their judgment policies. The policies were slightly less accurate than the actual assessments, but this situation is not surprising given that some part of the judgment policy may not be captured by a linear model (Cooksey, 1996). Therefore our first hypothesis was supported.

The judgment policies of the clinicians were not a straightforward computation of the Apgar score. The assessments were derived mostly from the heart rate sign and less from the other signs. The fact that the policies were highly accurate, despite not being made via the Apgar computation, can be explained with Brunswik's probabilistic functionalism. The cues that represent the state of the mannequin may have interdependencies (as with real neonates) and therefore cues can be partly interchangeable. Brunswik (1952) used the term *vicarious functioning* to describe this phenomenon and the term *vicarious mediation* to describe subjects' ability (consciously or unconsciously) to accurately capture and interpret these interdependent cues. From a physiological perspective, the clinical signs depend on blood circulation and cardiac activity and therefore a judgment policy with higher relative importance for the heart rate sign may accurately capture the 'real' clinical state of the mannequin.

When assessing the mannequin's illness severity, clinicians with more accurate policies tended to associate lower importance to the oxyhemoglobin saturation and higher importance to the muscle tone than did clinicians with less accurate policies. At this point we do not know if this finding is relevant only to judgments that were made in the simulated environment or whether it extends to assessments in actual practice. Altogether, we found that clinicians' assessments were probably largely

holistic and did not follow the Apgar computation. Therefore our second hypothesis was supported.

The fact that there was no advantage to more accurate policies when the clinicians participated in hands-on simulated resuscitation may indicate that the knowledge and ability to correctly interpret clinical signs were insufficient to ensure accurate assessments during hands-on simulated resuscitations. These results may indicate that during the hands-on condition clinicians' accuracy was differentiated by skills that were used to compensate for high task demand. This finding confirms our third hypothesis.

This study is relevant for training in healthcare in two ways. First, if clinicians are to make accurate judgments, their judgment policies should be accurate. With appropriate training and feedback, judgment policies can be adjusted (Cooksey, 1996).

Second, hands-on resuscitation usually involves high physical and mental workload. The hands-on scenarios in our study were designed to mimic real resuscitations and it is possible that clinicians were highly loaded when performing these scenarios. If clinicians used their policies while unable to monitor the mannequin's clinical signs effectively, this may have yielded less accurate judgments. To increase clinicians' monitoring capacity, training may need to focus on skills that enhance teamwork by encouraging clinicians to share information effectively, make clear role allocations, and distribute tasks more evenly among team members.

This investigation focused on assessment of the illness severity of neonates through the Apgar score. The applicability of the proposed approach to other clinical areas in which effective teamwork is critical should be tested in future studies. Future studies should also test the relation between policies captured in the simulated environment and judgments made in actual practice.

## CONCLUSION

Brunswik's theory and judgment analysis methods were used for modeling and testing clinicians' assessments. This solid theoretical framework and its associated methods have proven to be useful in a wide variety of studies, including the present one. The approach we have presented can be used to make training interventions more effective and clinical assessments more accurate. Consequently, clinical procedures may become more appropriate and patient outcomes may improve.

## ACKNOWLEDGMENTS

This paper was written as part of Izhak Nadler's PhD studies at The University of Queensland while he held Endeavour IPRS and UQILAS scholarships. Nadler's research activities were supported by grants from the Mater Mothers' Research Centre (research grant 1636), in kind contributions from Mater Health Services Brisbane Ltd., Australia and by an unrestricted donation from Laerdal Inc. We 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. We thank all the clinicians from the Mater Mothers' Hospital who volunteered to participate in the experiment. We gratefully acknowledge the advice and support of Ray Cooksey with respect to judgment analysis, but any errors are the authors' responsibility.

## REFERENCES

Apgar, V. (1953). A proposal for a new method of evaluation of the newborn infant. *Current researches in anesthesia & analgesia, 32*(4), 260-267.

Baker, P. N. (2006). *Obstetrics by ten teachers* (18th edition). London: Hodder Arnold.

Beckstead, J. W., & Stamp, K. D. (2007). Understanding how nurse practitioners estimate patients' risk for coronary heart disease: a judgment analysis. *Journal of Advanced Nursing, 60*, 436-446.

Brunswik, E. (1952). *The conceptual framework of psychology*. Chicago: University of Chicago Press.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments.* (2nd ed.). Berekeley CA: University of California.

Cooksey, R. W. (1996). *Judgment analysis: Theory, methods and applications*. San Diego: Academic Press.

Greenland, P., Bonow, R. O., Brundage, B. H., Budoff, M. J., Eisenberg, M. J., Grundy, S. M., et al. (2007). Clinical expert consensus document on coronary artery calcium scoring by computed tomography in global cardiovascular risk assessment and in evaluation of patients with chest pain. *Journal of the American College of Cardiology, 49*(3), 378-402.

Hammond, K. R., & Stewart, T. R. (2001). *The essential Brunswik: Beginnings, explications, applications*. New York: Oxford University Press.

Kirlik, A. (2010). Brunswikian theory and method as a foundation for simulation-based research on clinical judgment. *Simulation in Healthcare, 5*(5), 255-259.

Lowdermilk, D. L., & Perry, S. E. (2006). *Maternity nursing* (7th ed.). St. Louis, Mo.: Mosby Elsevier.

Murray, S. S., & McKinney, E. S. (2010). *Foundations of maternal-newborn and women's health nursing* (5th ed.). St. Louis, MO: Saunders Elsevier.

Nadler, I., Liley, H. G., & Sanderson, P. M. (2010). Clinicians can accurately assign Apgar scores to video recordings of simulated neonatal resuscitations. *Simulation in Healthcare, 5*(4), 204-212.

Nadler, I., Sanderson, P. M., & Liley, H. G. (in press). The accuracy of clinical assessments as a measure for teamwork effectiveness. *Simulation in Healthcare*.

Thompson, C., Bucknall, T., Estabrookes, C. A., Hutchinson, A., Fraser, K., de Vos, R., et al. (2009). Nurses' critical event risk assessments: a judgement analysis. *Journal of Clinical Nursing, 18*(4), 601-612.

Wigton, R. S. (2008). What do the theories of Egon Brunswik have to say to medical education? *Advances in Health Sciences Education*, 13(1), 109-121.