

# THE UNIVERSITY OF WARWICK

**Original citation:**

Paskins, Z., Kirkcaldy, J., Allen, M., Macdougall, C., Fraser, I. and Peile, E. (2010). Design, validation and dissemination of an undergraduate assessment tool using SimMan® in simulated medical emergencies. *Medical Teacher*, 32(1), pp. e12–e17.

**Permanent WRAP url:**

<http://wrap.warwick.ac.uk/6543>

**Copyright and reuse:**

The Warwick Research Archive Portal (WRAP) makes this work of researchers of the University of Warwick available open access under the following conditions. Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.


**Publisher's statement:**

<http://dx.doi.org/10.3109/01421590903199643>

**A note on versions:**

The version presented here may differ from the published version or, version of record, if you wish to cite this item you are advised to consult the publisher's version. Please see the 'permanent WRAP url' above for details on accessing the published version and note that access may require a subscription.

For more information, please contact the WRAP Team at: [wrap@warwick.ac.uk](mailto:wrap@warwick.ac.uk)

warwick**publications**wrap  
  
highlight your research

<http://go.warwick.ac.uk/lib-publications>

**Title**

**Design, validation and dissemination of an undergraduate assessment tool using SimMan® in **simulated** medical emergencies**

**Short Title**

SimMan assessment tool: design and validation

**Authors**

Zoe Paskins  
Jo Kirkcaldy  
Maggie Allen  
Colin Macdougall  
Ian Fraser  
Ed Peile

**Institutions**

University Hospitals of Coventry and Warwickshire NHS Trust, Coventry, UK  
Warwick Medical School, Gibbet Hill Lane, Coventry, UK

**Corresponding author**

Zoë Paskins  
SpR in Rheumatology  
Haywood Hospital  
High Lane  
Burslem  
Stoke on Trent  
ST6 7AG  
07813 304868  
Fax 01782 556306

## **Abstract**

### Background

Increasingly, medical students are being taught acute medicine using whole body simulator manikins.

### Aims

We aimed to design, validate and make widely available two simple assessment tools to be used with Laerdal SimMan® for final year students.

### Methods

We designed two scenarios with criterion-based checklists focused on assessment and management of two medical emergencies. Members of faculty critiqued the assessments for face validity and checklists revised. We assessed three groups of different experience levels: Foundation Year 2 doctors, 3<sup>rd</sup> and final year medical students. Differences between groups were analysed, internal consistency and interrater reliability calculated. A generalizability analysis was conducted using scenario and rater as facets in design.

### Results

A maximum of two items were removed from either checklist following the initial survey. Significantly different scores for three groups of experience for both scenarios were reported ( $p < 0.001$ ). Interrater reliability was excellent ( $r > 0.90$ ). Internal consistency was poor ( $\alpha < 0.5$ ). Generalisability study results **suggest** that four cases would provide reliable discrimination between final year students.

## Conclusions

These assessments proved easy to administer and we have gone some way to demonstrating construct validity and reliability. We have made the material available on a simulator website to enable others to reproduce these assessments.

## **Practice Points**

- Acute care skills are gaining more importance at medical undergraduate level, but are not always assessed
- This study demonstrates feasibility of two checklist assessments using a widely used whole-body simulator, and contributes to establishing validity and reliability of these assessments.
- We have made all the assessment material available on the web for others' use, to avoid duplication of work.

## **Introduction**

Since Tomorrow's Doctors was published, undergraduate medical education has changed to place more emphasis on clinical skills and to increase preparedness for the junior doctor's role (General Medical Council 2003). The Acute Care Undergraduate Teaching (ACUTE) Initiative was published in 2005 (Perkins, Barret et al. 2005), in response to a number of publications raising concerns about the care of acutely ill patients (McQuillan, Pilkington et al. 1998; Franklin and Mathew 2002; Hodgetts, Kenward et al. 2002; Cullinane, Findlay et al. 2005). This report details competencies in the care of acutely ill patients which the group suggests should be integrated into undergraduate curricula.

Undergraduate acute care skills are most commonly assessed by "paper simulation"; however, written examinations are more likely to test knowledge alone rather than the complex integration of applied knowledge with clinical skills and problem solving ability. Simulator manikins can be used for observation-based competence assessments, to enable a higher level of Miller's pyramid to be assessed: "shows how" (Miller 1990).

Simulator manikins are being increasingly used in undergraduate education (Bradley 2006). These manikins vary in sophistication and technical detail (ranging from low to high fidelity) but most are able to reproduce the haemodynamics of the critically ill patient, making them ideally suited for teaching acute care skills. Simulators are also ideally placed for use in evaluating students' acute care skills; the environment is safe, assessments can be easily standardised and importantly, the assessment setting may have more authenticity than traditional assessment methods (Schuwirth and Van der

Vleuten 2003). Furthermore, with increasing student numbers the need to develop assessment methods that do not involve patients is great (Maran and Glavin 2003; Bradley 2006).

There is much in the literature concerning the reliability and validity of assessment tools using simulators, mostly in the anaesthetic field. A 2001 review of 13 papers reporting design of assessment tools for doctors using high-fidelity simulators was critical of the reliability and validity evaluations made (Byrne and Greaves 2001). Since this review, further studies using high-fidelity simulators have been published reporting reliability and validity of checklist assessments in anaesthetics and medical emergencies (Morgan and Cleave Hogg 2000b; 2001b; Murray, Boulet et al. 2002; Boulet, Murray et al. 2003; Gordon, Tancredi et al. 2003; Morgan, Cleave Hogg et al. 2004). Previous studies regarding undergraduate assessments in this area have reported that checklist assessments **are associated with** high interrater reliability and **have demonstrable** construct validity, determined by assessing differing experience levels (Devitt, Kurrek et al. 1998; Morgan and Cleave Hogg 2000a; Devitt, Kurrek et al. 2001; Morgan, Cleave Hogg et al. 2001b; Murray, Boulet et al. 2002; Boulet, Murray et al. 2003; Morgan, Cleave Hogg et al. 2004). Convergent validity has not been established, in that simulator assessment results do not correlate well with other assessments e.g. written (Morgan, Cleave Hogg et al. 2001b). It is **possible** that this **is due** to written assessments testing different constructs to that of simulator assessments.

More recently the Laerdal SimMan® has become available which is “moderate or medium fidelity”, lower in cost and, according to the manufacturers has 90%

of the market share in the UK. The SimMan® has many similar features to high-fidelity models, but may be less suited to certain scenarios e.g. neurological emergencies, since its pupils are non-reactive.

Designing and validating assessment tools is time-consuming and a lengthy process. Previous studies in this area, including one using SimMan® (Weller, Robinson et al. 2004) are extremely useful to use as a framework on which to base further tool evaluation, however no previous work to the authors' knowledge have made all the material (including software programmes) available for others to reproduce the assessments and therefore avoid duplicating the validation process for their own assessments.

### Aim

The primary aim of this study was to develop a robust formative assessment that could be used to assess the acute care skills of final year medical students at the end of an Emergency Medicine attachment using the widely available SimMan®. The assessment tool was designed to operate with limited resources, so that it was feasible and practical to deliver to a reasonable number of students (on average, 15) rotating every three weeks. Our secondary aim was to disseminate the results and tools, including checklists and pre-programmed software, so that other centres could easily make use of our assessment material.



## **Methods**

### **Simulator and setting**

We used the Laerdal SimMan®. This medium-fidelity simulator is a life-size manikin that breathes, talks, has palpable pulses, audible chest, heart and bowel sounds, and is connected to a monitor for displaying oxygen saturations, ECG trace, pulse rate and blood pressure. The manikin is connected to a computer, and the assessment scenarios were pre-programmed for consistency; each time we used a scenario the parameters (pulse, breath sounds, oxygen saturation etc.) were the same. Furthermore, the software enables pre-programmed standard responses to student actions e.g. administering oxygen.

The SimMan® is located in a clinical skills laboratory with appropriate “props” such as oxygen masks, cannulation equipment, and fluids. In addition, for the scenarios used, there were standardised ECGs and Arterial Blood Gas results for the students, if requested. Participants were given an identical structured introduction to SimMan® prior to the assessment. Two assessors (ZP and JK) were present for all assessments. One operated the software and provided the voice of SimMan® for history points. The other gave each student an introduction prior to the assessments, standard prompts during the assessment if necessary and also acted as an assistant able to perform clinical observations and cannulate.

### **Instrument**

We designed two scenarios based on the assessment and management of acute coronary syndrome (ACS) and acute severe asthma (AA), lasting approximately 10 minutes each; **these emergencies were chosen as they were felt to be easily simulated using the Laerdal SimMan®.** A criterion based checklist was developed for each scenario. The items in each checklist included aspects of Airway, Breathing and Circulation assessment (ABC), eliciting pertinent history and examination findings, requesting and interpreting investigations and initiating basic management steps. We designed the checklist content to correspond with the relevant objectives of the Medical School Curriculum and also the ACUTE initiative, **for the two scenarios chosen** (Perkins, Barret et al. 2005). **The Trust's Clinical** Ethics Committee deemed that formal ethical approval was not necessary.

To establish face validity, checklists and scenarios were circulated to 22 consultants involved in undergraduate teaching and emergency medicine, who were asked to indicate whether each task was appropriate for final year undergraduates. Consultants were asked to rank the tasks in order of importance on a 3-point Likert scale; the mode of these answers (score 1 – 3) was taken as a score for each task and used to weight each component in order of importance. If more than 20% of the consultants felt a task was inappropriate, it was removed.

The checklist included one aspect of timed assessment (time taken to assess ABC). The checklist was completed independently by both assessors (ZP, JK) for all candidates.

A pilot was run with 12 final year students resulting in a number of minor changes: clinical information in the scenarios was changed slightly as some details were ambiguous; checklists were modified to include standard prompts the examiner should say if a task was not performed e.g. “the oxygen saturations are still low”; and marking guidelines were produced for clarification of items where scoring had been troublesome e.g. medications for which students were expected to know both dose and route in order to score marks. In addition, the assessment was stopped if not completed after 10 minutes, since most candidates in the pilot had completed the test in this time. This was primarily to increase the feasibility of using the tool, but also acted to prevent an extremely slow candidate scoring the same as an efficient one. The scoring system was changed to reflect the use of prompts, so the student could only score half marks for a correct item, if prompted.

## **Participants**

To assess construct validity, both assessment tools were administered to three groups of volunteers with different experience levels; 20 third year (graduate entry, 4 year course) medical students, 18 final year students and 24 Foundation Year 2 doctors (FY2). All medical students involved had no previous exposure to SimMan®. Participants received both assessments on the same day; the order was alternated so that 50% of each group received the AA scenario first. All participants were given detailed feedback after their performance, by an assessor or an independent observer; either on the same day, or three days after the assessment. Anonymity was subsequently maintained using numbers to identify the individuals.

## **Analysis**

We assessed interrater reliability using intraclass correlation. Internal consistency of both checklists was measured using Cronbach's alpha. The difference between three groups of experience level was calculated using one-way ANOVA. SPSS Versions 12.0 and 15.0 were used for statistical analysis. A generalizability analysis was conducted using GENOVA Version 3.1.

## **Results**

1 item was removed from the ACS checklist and 2 items from the AA checklist following the assessment of face validity. The final checklists used (after consultant survey and pilots) are available at [http://simulation.laerdal.com/forum/files/folders/user\\_scenarios/default.aspx](http://simulation.laerdal.com/forum/files/folders/user_scenarios/default.aspx).

The mean scores for the ACS assessment were 25.1, 36.2, and 47.9 for the third year, final years and FY2 respectively (Figure 1), out of a maximum score of 67. The mean score for the AA assessment was 28.6, 39.1 and 49.7 respectively, out of a maximum score of 72 (Figure 2). The difference between all groups for both assessments was statistically significant ( $p < 0.001$ ).

There was no significant difference between the sex distribution of the three groups ( $p = 0.495$ ). Two of the FY2s had had brief exposure to SimMan® before. If these 2 individuals results were discounted as a possible source of bias, the mean of the FY2 scores were 47.76 (ACS) and 50.1 (AA), which remain significantly different to the other groups ( $p < 0.001$ ).

The reliability measures are detailed in Table 1. Deletion of any item on either checklist did not result in any substantial improvement in Cronbach's alpha, and therefore no items were removed.

A generalizability analysis was conducted using rater and case (scenario) as facets in the design (a two facet crossed design). The three groups of experience level were analysed separately, to minimize examinee variation. The summary of these results are tabulated in Table 2. The variance

components represent error variance, and their magnitudes reveal the importance of the various sources of error (Mushquash and O'Connor 2006). Using the data from the generalizability analysis, the G study, one can conduct a Decision or D study to evaluate the effectiveness of alternative designs with differing numbers of facets. An example is shown in Table 3.

## **Discussion**

We have produced two instruments which have demonstrable interrater reliability and face validity, and our findings of increased scores with experience are in support of construct validity. We have made available the scenarios, assessment forms, history points, marking guidance and “props” (ABGs, ECGs) on the Laerdal Simulation User Network, [http://simulation.laerdal.com/forum/files/folders/user\\_scenarios/default.aspx](http://simulation.laerdal.com/forum/files/folders/user_scenarios/default.aspx).

Sharing assessment tools and scenarios among educators permits further assessment of reliability and validity and encourages standardisation (Bond and Spillane 2002). **Although there are ever increasing resources available on the web for use with simulators, this is the first study to our knowledge for the Laerdal SimMan® that has reported the validation process and made available all the material necessary to reproduce the assessments, particularly the programmed scenarios.**

The assessments are feasible and easy to administer, requiring two members of staff, and we found it possible to individually assess a group of 15 students in two and a half hours (one assessment scenario). There is some rationale for time-limitation in checklist assessment of scenarios that in real-life require both rapid clinical reasoning and performance of clinical skills: experts perform better on speeded up sensorimotor tasks where attention to execution is limited, in contrast to novices whose performance improves with additional time to attend to detail (Beilock, Bertenthal et al. 2004).

The generalizability analysis shows that the largest variance component was for examinees, which is to be expected, and not a source of error (Mushquash

and O'Connor 2006). The next largest variance component across all three groups, was examinee x case, which indicates the rank ordering of examinees differed across the two cases. The G coefficients reflect the reliability of the scores across raters and cases and are reasonably close to the conventional threshold of 0.80, for 3<sup>rd</sup> and 4<sup>th</sup> year medical students. The G coefficients are based on relative, rather than absolute decisions; if absolute decisions are required, the reliability will be lower. Although this study contains small numbers for this type of analysis, the D study demonstrates that four cases with one rater would be desirable for a G co-efficient of  $\geq 0.8$  for final year medical students, for which the tool was designed. Boulet et al (2003) found that student performance did not generalize well from one case to another, supporting the notion that multiple cases are necessary. However, technical limitations of the SimMan® may prevent the whole range of medical emergencies in the acute care curriculum being sampled e.g. neurological. Even when further cases have been designed and evaluated, it would still be unsafe to assume that achieving a G co-efficient of  $\geq 0.8$  across all the cases would ensure that students had been robustly assessed on their ability to manage any emergency.

Students have valued exposure of deficits in their ability to assess and treat acutely ill patients, and welcomed the idea of using SimMan in end-of-year assessments (MacDowall 2006). In aiming to further the adoption of these instruments summatively, our study has a number of limitations. The first is the sampling bias of only using two scenarios, as discussed above. Secondly, the range of domains assessed within each scenario could have been expanded: for instance, we made no assessment of communication skills; while we



acknowledge that simulation exercises are hugely important in teaching communication skills, we decided not to assess these in the interest of keeping our instrument simple, and because communication skills were not explicit in the objectives of our assessment. The nature of a checklist assessment prohibits the inclusion of complex cases which may also be detrimental to content validity (Schuwirth and Van der Vleuten 2003). However, balance must be sought between authenticity and feasibility and the primary aim of this study was to produce a tool that is easy to administer. **Standardisation across cases could have been improved by having a set time before issuing a prompt for each item.**

A further barrier to summative implementation may be inferred from the low 'Cronbach's alpha' **measure of internal consistency** achieved in this study and **others** (Devitt, Kurrek et al. 1998; Morgan, Cleave Hogg et al. 2004). Values of  $> 0.7$  are desirable for high stakes assessment, **and most values were below 0.5 in this study; furthermore the reported values are likely to have** been adversely affected by the large number of missing values for items, particularly in the 3rd year students.

However, Cronbach's alpha should be interpreted with caution in checklist assessments where items are not random parallel, i.e. not randomly sampled from the total possible number of items, and not truly independent (Cronbach 2004). In our assessment, there may have been more items to represent ABC assessment, for example than other domains, and performance in one item may have affected performance in another, and therefore these assumptions have not been met. Furthermore, Cronbach indicated that alpha should not be

used if a time limit has been set to a test so that part of the scores may equal zero (due to running out of time) (2004). In our study only one of the 4<sup>th</sup> years (with no FY2s and eight of the 3<sup>rd</sup> years) failed to complete one or both tests. Omission of the 4<sup>th</sup> year student's result who failed to complete increased Cronbach's alpha slightly ( $\alpha = 0.349$ ). Murray et al have been critical of previous researchers placing too much emphasis on item-item correlations in the assessment of internal consistency; to remove a test item based on statistical results without considering the clinical significance of the item may be sacrificing validity for reliability (2002).

Our interrater reliability was found to be excellent, and comparable with other studies (Morgan and Cleave Hogg 2000b; 2001b; 2004). This was probably influenced by our action after the pilot study in producing prompt sheets for markers; similar observations have been noted in previous work (Murray, Boulet et al. 2002). The exclusion of behavioural aspects such as communication skills which are likely to be difficult to measure is also likely to have increased measured interrater reliability. Interrater reliability in this study is based on the results of two authors who were clearly intrinsically involved in the scenario design. We have, however, measured interrater reliability between one author, and a member of faculty not involved in the study on a small group of students (11 final year students, AA checklist) with  $r = 0.890$ ; this suggests that the interrater reliability could be generalised to other assessors. We could have tested interrater reliability amongst "non-experts". Boulet et al found little difference between nurse clinicians and faculty members in rating students on criterion checklists (2003). An advantage of using experts is that a global judgement of overall performance can be incorporated into the scoring strategy.

This has been found to have equivalent reliability to checklists (Morgan, Cleave Hogg et al. 2001a) and is suggested may yield more valid results (Boulet, Murray et al. 2003). Again, we elected not to do this so that non-experts could rate, although this still needs evaluation.

Face validity was established among 22 of the teaching faculty. We could have also surveyed the students' views of the assessment; other studies report positive evaluations (Morgan and Cleave Hogg 2000b; Weller, Robinson et al. 2004). Face validity is often overlooked but is intrinsically linked to a student's motivation in taking a test and therefore of importance in assessment design (Guilford 1954). Clearly the assessment itself should not be considered in isolation, and it is the feedback which the student receives after the assessment which is key. Further formal evaluation of this feedback would add value to the design of the assessment process.

With the proviso that the raters were not blinded to the level of experience of the participants, the scenarios and checklists can be said to measure a construct that increases with medical experience and we can infer that acute care skills would also improve with experience. However, we cannot state from this study alone that we have measured the construct of acute care skills without further work such as confirmatory factor analysis. Clearly fully evaluating construct validity in the context of emergency medicine is problematic, not least due to the difficulty in defining the construct.

## **Conclusion**

In conclusion, we have demonstrated the reliability and validity of two user-friendly assessment tools using the Laerdal SimMan®. With the dissemination of these results and materials necessary to reproduce the assessments, other medical schools may adopt these instruments for formative use. Further work to expand the range of scenarios may enable these assessments to be incorporated in summative examinations, and our G study results suggest four scenarios would provide a robust measure, although this would need further evaluation. Perhaps more importantly, the question still remains as to whether simulator training makes better doctors or not; evaluating the predictive validity of simulator use remains somewhat of a holy grail for medical education researchers.

## **Acknowledgements**

The authors would like to acknowledge the help and assistance of Dr Lois Brand and Dr Denis Lindo with data collection and Professor David Wall for his help with statistical analyses.

## **Declaration of interest**

*The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.*

## **List of Tables and Figures**

Figure 1: Acute Coronary Syndrome (ACS) Score across three groups of experience

Figure 2: Acute Asthma (AA) Score across three groups of experience

Table 1: Reliability results: interrater reliability and internal consistency

Table 2: Variance Component Matrix from Generalisability Analysis

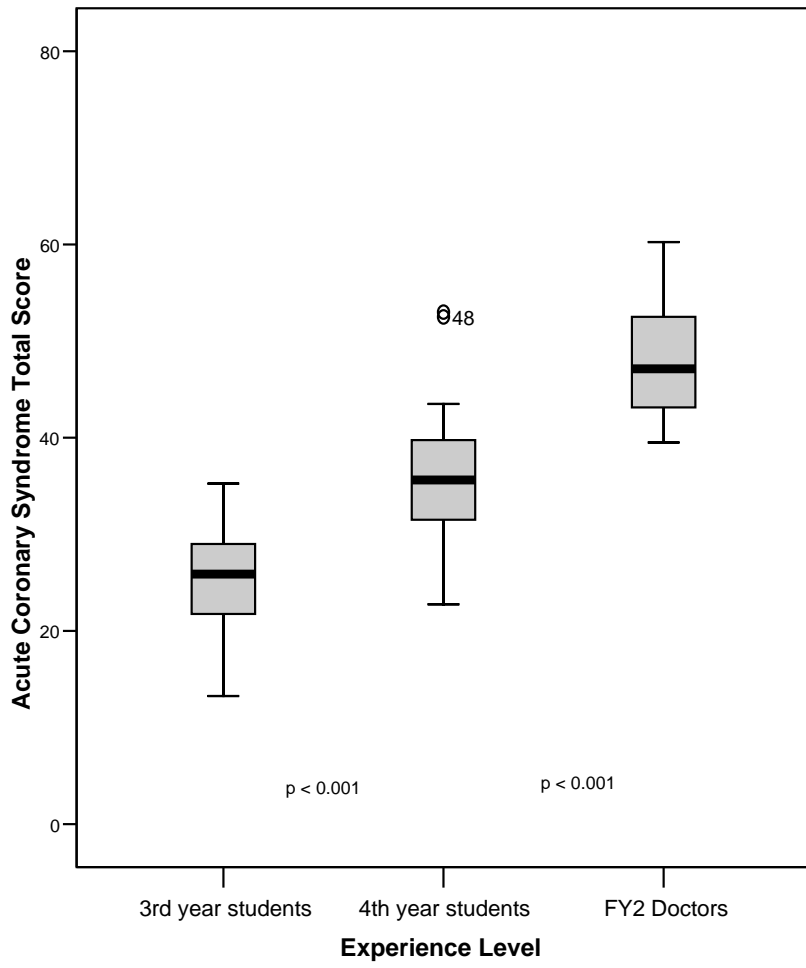
Column A = Variance components (proportion total variance)

Column B = Standard Error of the variance component for mean scores

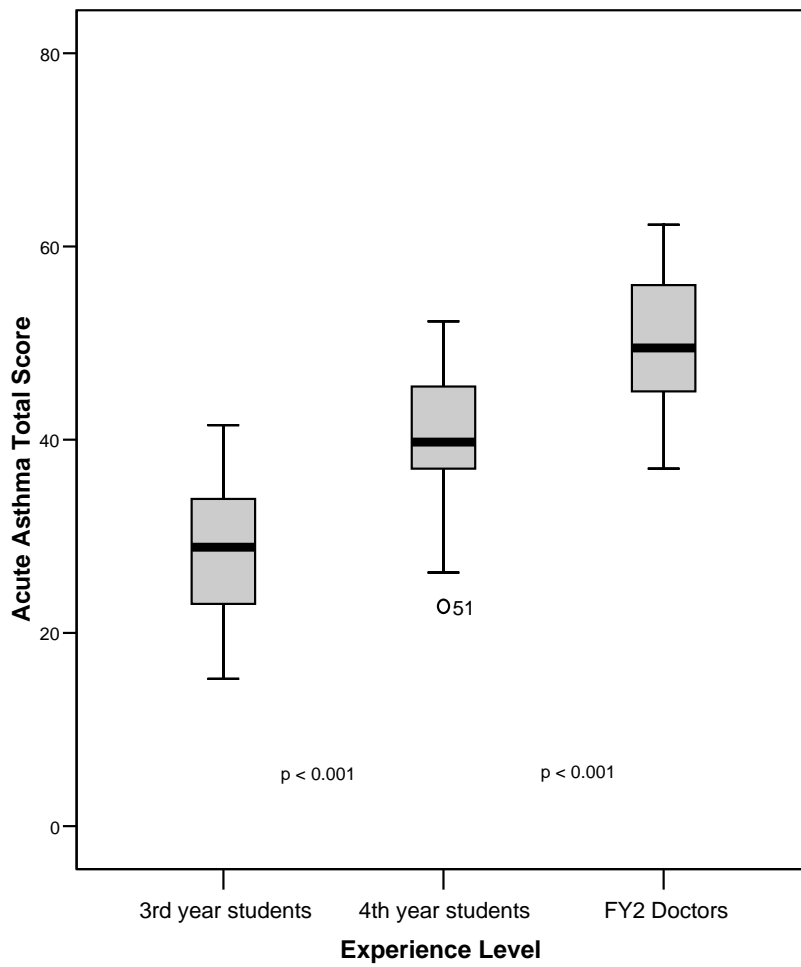
Table 3: D study to examine effect on variance of more cases and number of assessors for each study group of candidates

Nb. Shaded boxes indicate an adequate number of cases and raters to achieve acceptable reliability at G-coefficient of variance 0.8.

**Figure 1**



**Figure 2**



**Table 1**

Experience Group	Interrater reliability		Cronbach's Alpha	
	ACS	AA	ACS	AA
<b>3<sup>rd</sup> Years</b>	0.917	0.923	0.197	0.335
<b>4<sup>th</sup> Years</b>	0.936	0.934	0.334	0.273
<b>FY2 Doctors</b>	0.928	0.966	0.494	0.453

**Table 2**

	3 <sup>rd</sup> Year Students		4 <sup>th</sup> Year Students		FY2 Doctors		Total group	
	A	B	A	B	A	B	A	B
Examinee	23.421 (0.477)	10.533	34.716 (0.516)	16.383	18.478 (0.424)	8.95	110.082 (0.802)	21.715
Case	5.466 (0.111)	2.594	2.807 (0.042)	1.158	0.726 (0.017)	0.699	3.292 (0.024)	1.490
Rater	0.000 (0.000)	0.028	0.107 (0.002)	0.066	0.807 (0.019)	0.341	0.020 (0.000)	0.024
Examinee x case	16.615 (0.338)	2.759	25.311 (0.376)	2.938	21.738 (0.499)	3.203	20.690 (0.151)	1.956
Examinee x rater	1.156 (0.024)	0.413	0.705 (0.010)	0.499	0.000 (0.000)	0.137	0.667 (0.005)	0.201
Case x rater	0.000 (0.000)	0.022	0.000 (0.000)	0.011	0.023 (0.001)	0.021	0.008 (0.000)	0.010
Examinee x case x rater	2.447 (0.050)	0.188	3.584 (0.053)	0.194	1.789 (0.041)	0.127	2.458 (0.018)	0.109
<b>G Coefficient</b>	0.711		0.714		0.620		0.907	

**Table 3**

No of cases	3 <sup>rd</sup> Year Students		4 <sup>th</sup> Year Students		FY2 Doctors	
	1 Rater	2 Raters	1 Rater	2 Raters	1 Rater	2 Raters
2	0.687	0.711	0.696	0.714	0.608	0.617
3	0.757	0.782	0.771	0.787	0.699	0.707
4	0.798	0.823	0.814	0.830	0.756	0.763
5	0.825	0.849	0.843	0.857	0.795	0.801
6	0.844	0.868	0.863	0.877	0.823	0.828



## **References**

- Beilock, S. L., Bertenthal, B. I., et al. (2004). Haste does not always make waste: Expertise, direction of attention, and speed versus accuracy in performing psychomotor skills. *Psychonomic Bulletin & Review*,(11), pp. 373-9.
- Bond, W. F. and Spillane, L. (2002). The use of simulation for emergency medicine resident assessment. *Academic Emergency Medicine*, **9**(11), pp. 1295-9.
- Boulet, J. R., Murray, D., et al. (2003). Reliability and validity of a simulation-based acute care skills assessment for medical students and residents. *Anesthesiology*, **99**(6), pp. 1270-80.
- Bradley, P. (2006). The history of simulation in medical education and possible future directions. *Medical Education*, **40**(3), pp. 254-62.
- Byrne, A. J. and Greaves, J. D. (2001). Assessment instruments used during anaesthetic simulation: review of published studies. *British Journal of Anaesthesia*, **86**(3), pp. 445-50.
- Cronbach, L. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and psychological measurement*, **64**(3), pp. 391-418.
- Cullinane, M., Findlay, G., et al. (2005). "National Confidential Enquiry into Patient Outcome and Death. An Acute Problem." Available from <http://www.ncepod.org.uk/2005report/>.
- Devitt, J. H., Kurrek, M. M., et al. (2001). The validity of performance assessments using simulation. *Anesthesiology*, **95**(1), pp. 36-42.
- Devitt, J. H., Kurrek, M. M., et al. (1998). Testing internal consistency and construct validity during evaluation of performance in a patient simulator. *Anesthesia and analgesia*, **86**(6), pp. 1160-4.
- Franklin, C. and Mathew, J. (2002). Developing strategies to prevent in-hospital cardiac arrest: analyzing responses of physicians and nurses in the hours before the event. *Critical Care Medicine*,(22), pp. 244-7.

- General Medical Council (2003). *Tomorrow's Doctors*. London: General Medical Council.
- Gordon, J. A., Tancredi, D. N., et al. (2003). Assessment of a clinical performance evaluation tool for use in a simulator-based testing environment: a pilot study. *Academic Medicine*, **78**(10 Suppl), pp. S45-7.
- Guilford, J. P. (1954). *Psychometric methods*. 2nd ed. New York: McGraw-Hill Book Co.
- Hodgetts, T. J., Kenward, G., et al. (2002). Incidence, location and reasons for avoidable in-hospital cardiac arrest in a district general hospital. *Resuscitation*,(54), pp. 115-23.
- MacDowall, J. (2006). The assessment and treatment of the acutely ill patient - the role of the patient simulator as a teaching tool in the undergraduate programme. *Medical Teacher*, **28**(4), pp. 326-9.
- Maran, N. J. and Glavin, R. J. (2003). Low- to high-fidelity simulation - a continuum of medical education? *Medical Education*, **37**(Suppl 1), pp. 22-8.
- McQuillan, P., Pilkington, S., et al. (1998). Confidential inquiry into quality of care before admission to intensive care. *British Medical Journal*,(316), pp. 1853-8.
- Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine*, **65**(9), pp. S63-7.
- Morgan, P. J. and Cleave Hogg, D. (2000a). A Canadian simulation experience: faculty and student opinions of a performance evaluation study. *British Journal of Anaesthesia*, **85**(5), pp. 779-81.
- Morgan, P. J. and Cleave Hogg, D. (2000b). Evaluation of medical students' performance using the anaesthesia simulator. *Medical Education*, **34**(1), pp. 42-5.
- Morgan, P. J., Cleave Hogg, D., et al. (2004). High-fidelity patient simulation: validation of performance checklists. *British Journal of Anaesthesia*, **92**(3), pp. 388-92.

- Morgan, P. J., Cleave Hogg, D., et al. (2001a). A comparison of global ratings and checklist scores from an undergraduate assessment using an anesthesia simulator. *Academic Medicine*, **76**(10), pp. 1053-5.
- Morgan, P. J., Cleave Hogg, D. M., et al. (2001b). Validity and reliability of undergraduate performance assessments in an anesthesia simulator. *Canadian Journal of Anaesthesia*, **48**(3), pp. 225-33.
- Murray, D., Boulet, J., et al. (2002). An acute care skills evaluation for graduating medical students: a pilot study using clinical simulation. *Medical Education*, **36**(9), pp. 833-41.
- Mushquash, C. and O'Connor, B. P. (2006). SPSS and SAS programs for generalizability theory analyses. *Behavior Research Methods*, **38**(3), pp. 542-7.
- Perkins, G. D., Barret, H., et al. (2005). The Acute Care Undergraduate TEaching (ACUTE) Initiative: consensus development of core competencies in acute care for undergraduates in the United Kingdom. *Intensive Care Medicine*, **31**(12), pp. 1627-33.
- Schuwirth, L. W. T. and Van der Vleuten, C. P. M. (2003). The use of clinical simulations in assessment. *Medical Education*, **37**(Suppl 1), pp. 65-71.
- Weller, J., Robinson, B., et al. (2004). Simulation-based training to improve acute care skills in medical undergraduates. *The New Zealand Medical Journal*, **117**(1204), pp. U1119.

