PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	The effect of a proficiency-based progression simulation programme on clinical communication for the deteriorating patient: A randomised controlled trial.
AUTHORS	Breen, Dorothy; O'Brien, Sinead; McCarthy, Nora; Gallagher, Anthony; Walshe, Nuala

VERSION 1 - REVIEW

REVIEWER	Alisa Khan, MD, MPH Staff Physician, Boston Children's Hospital Instructor in Pediatrics, Harvard Medical School USA
REVIEW RETURNED	26-Sep-2018

GENERAL COMMENTS	Major: • Handovers. Handovers are mentioned in the title and throughout the paper. However, it appears that the authors are not talking about handovers/handoffs in the traditional sense but rather discussions between providers about acute patient deteriorations. Consider clarifying or removing the use of the term handover. • ISBAR. ISBAR is used in a manner that is a bit confusing. I would argue that effectively handling a deteriorating patient, rather than whether ISBAR is used or not, is the actual question of importance. Thus, instead of saying “approach to clinical handover (ISBAR)” in the abstract, the authors could say “approach to a deteriorating patient,” moving ISBAR to the abstract methods. Similarly, in the main text methods, instead of saying “effectiveness on ISBAR performance,” they could say “effectiveness of communication in the context of a clinically deteriorating patient.” My impression is that ISBAR is a means to an ends and not the main purpose of the intervention. • Benchmarks and metrics. Please provide the benchmark rubric or at the very least, examples of benchmarks and pre-defined metrics. It is important for the reader to have a better sense of what the benchmarks and 24-26 metrics per scenario are. The results would generally benefit from additional detail and findings surrounding the proficiency benchmarks and predefined metrics. • Scoring by partners. Please justify the fact that partners scored each other’s phone calls during training. Perhaps this was due to feasibility or because the rubric was entirely objective (again, including the rubric they used would help). However, might this have lessened the efficacy of the performance-based progression as partners might be more apt to “graduate” their partner to the next level? Consider adding this to the limitations. • Defining terms. Please define ISBAR and proficiency-based progression in the abstract. Please also consider providing examples of proficiency benchmarks in the last paragraph of the

	introduction (“e.g., tying a surgical knot in X seconds”). Please also define more clearly the Health Services Executive in the introduction and Transfer of Training in the methods for readers unfamiliar with these terms. Minor: • Use of abbreviations. The abbreviations used for the 3 arms of the study can be confusing (particularly HSE) and made findings often difficult to follow. Consider eliminating them entirely or replacing with more intuitive abbreviations, like “e-learning (E)” instead of HSE. The S and PBS abbreviations are easier to follow, but could also consider alluding to the fact that these arms included e-learning as well. For instance, these abbreviations could be “e-learning with simulation (E+S)” vs. “e-learning with proficiency-based progression (E+PBS)” though I think S and PBS alone would be fine if HSE was replaced with E for example. Please also avoid using these abbreviations in Figure 3 (or define them so the figure stands alone). • Examples of simulation scenarios, steps, errors, and critical errors. Please consider providing a complete list of these, or at the very least, an example of each of these categories. • Please avoid stating there is >2 times the odds when the results are not statistically significant (e.g., page 17, lines 39-42; Figure 4) • Kappas. Interrater reliability was 85%, but what was the kappa? • Please provide p-values for Table 1 characteristics if possible. Editorial: • Page 16, lines 10-21 is redundant with page 15, lines 11-16. Could perhaps remove one of these sections.

REVIEWER	Liaw Sok Ying National University of Singapore
REVIEW RETURNED	05-Oct-2018

GENERAL COMMENTS	The authors have used rigorous methodology, 3-arm RCT and performance outcomes, to evaluate the educational interventions. The study method was very well described. However, the introduction and discussion sections need more-depth information. More justifications are required to support the implementation of a proficiency-based progression simulation over the standard simulation training. The pedagogical concepts underpinning these simulation approaches need to be described. These concepts should also be applied and discussed to justify the outcomes of the study. I would also suggest removing the unclear phrase "compared to standard training" from the title. I hope these comments will help you to improve the quality of the paper.

REVIEWER	Daryl Cheng The Hospital for Sick Children Toronto Canada
REVIEW RETURNED	07-Oct-2018

GENERAL COMMENTS	This study provides an evaluation of proficiency based progression training for ISBAR based deteriorating patient score escalation communication. It is a well written manuscript which compares three arms - standard, simulation and PBP. It adds to current literature as evaluating PBP in a non-technical skill setting. Considerations for Authors - There is a lot to digest in tackling this topic, and I think this can be spelled out more clearly in the objectives. The assumption is that the authors are not addressing the validity of ISBAR and its use in this setting; and likewise not addressing effectiveness of other components of NEWS training (elearning, simulation delivery etc) - but only addressing PBP vs other types of delivery. If that is the case, it would help to make this more clearly evident within the background/objectives sections. One way to consider would be to signpost more clearly the concepts of ISBAR and NEWS, but then discuss about PBP as a concept in more detail. Likewise, expansion with the discussion and limitations sections about ISBAR and NEWS and their confounding effects on results needs to be included. Currently - the conclusion (?assumption) listed above is only reached if the reader is familiar with 1. ISBAR; 2. Early deterioration scores 3. PBP and simulation - which may make the manuscript less generalisable. 2. Page 9 line 22 – references are not displayed properly (superscript) 3. Power calculation is difficult to follow and is potentially flawed – if the numbers is based on previous technical skills testing in cardiology and surgery (very different scenarios and settings; and different levels of training), it may not be an accurate way to calculate an appropriate sample size. Although using this method sample size is appropriate, at a bigger picture glance this may be difficult to justify with n=30 per group chosen. it would be good to discuss this in limitations and in future studies a more robust and expanded methodology with a larger sample size chosen - this would make the study more generalisable, reproducible, and ultimately make the observed effect stronger.

REVIEWER	Ramesh Walpola Griffith University, Queensland, Australia
REVIEW RETURNED	12-Oct-2018

GENERAL COMMENTS	The paper was generally very well written and provides insight into a training method for an important skill in healthcare professional education. The paper follows the correct reporting conventions of an RCT and is appropriate for publication in BMJ Open. I only have

	minor changes to the article to improve readability. Specific comments: Please move the Aims to the introduction section of the paper and include more specific objectives. Please provide more information of the scoring process or provide a copy of the scoring form as an online appendix. Please provide more detail in your conclusion section.

REVIEWER	Dr. Zoë Hoare NWORTH CTU Bangor University
REVIEW RETURNED	29-Oct-2018

GENERAL COMMENTS	The manuscript describes the results from a single site trial for the delivery of a training technique. The study has been described adequately but there a few areas where the manuscript could have been improved. The randomisation process needs greater clarity about the exact methodology used. For the PBP group how long does the training session last for? It is indicated that participants were required to reach proficiency on all four cases and this repeated in a cyclical fashion but does not appear to place a time limit on these iterations. Although a time limit of 3.5 hours is mentioned for the simulation training. Is there a need within the analysis of the results to be able to assess the effect of any of the demographics? Effectiveness of the methodology would be berrer proved using a mutli-centre design accomodating for the facilitator effect. There is a large effect evident in this data but this is one undergraduate year in one centre with experienced facilitators who have been involved in the development of this work package. It would be necessary for conclusive evidence to widen this scope and generate more evidence for this promising intervention. How this could be implemented across an undergarduate training program is not considered here and would be worthy of mentioning - evidence of effect will only be bourne out if there is evidence of implemenation. Reference is made to a pilot study but no literature reference is given? Has this process data been published?

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

Thank you for your feedback

1. The term handover has been removed and replaced with communication.

2. The phrase "communication in the context of a clinically deteriorating patient" as recommended is used in both the abstract and the main text methods now. ISBAR moved to methods section.

3. Examples of a case and benchmark rubric given in table 2 and 3

4. Partners scored each other during training to maximise deliberate practice. The metrics were sufficiently objective as now given in example format in table 3. Furthermore candidates were required to demonstrate proficiency with the facilitator. If not proficient with the facilitator then candidates had to practice again with the partner. See methods, interventions (iii). Partners were used for two reasons so that practice was deliberate (rather than repeated) and to maximise the use of resources and training time. Demonstrating proficiency in training with the facilitator in training was used an additional step to verify performance. For the performance assessment digital recordings were used and reviewed by 2 independent and blind assessors.

5. ISBAR and proficiency based progression are now defined in the abstract. The term Health Service Executive now removed and replaced with Irish Health Service. Transfer of training now also defined in the statistics section.

Minor

6. The terms HSE, S and PBP have been replaced with E, E+S, and E+PBP as recommended. Figure 3 amended and terms defined in the legend.

7 Example of a scenario steps, errors and critical errors given table 2 and 3

8. "Please avoid stating there is >2 times the odds when the results are not statistically significant (e.g., page 17, lines 39-42; Figure 4) kappa"

The results section is simply reporting the results that were observed from the statistical analysis. We quite rightly reported that the Exponential of B = 2.04. This means that the simulation trained group were two times as likely to demonstrate the proficiency benchmark as the HSE trained group. We do however accept the point made by the reviewer that this looks a bit odd, i.e., 'two times more likely', but not statistically significant. We have therefore changed this section to read more clearly and qualified the 'two times more likely' statement.

"On logistic regression analysis (figure 4) it was found that in comparison to the HSE group, the S group were 2 times as likely to demonstrate proficiency (Ext (B) =2.04, 95% CI=0.31-13.28, p=0.46). This difference was in the direction of improved performance but the effect was not statistically significant probably because of the sample size used in this study. In contrast the PBP trained group were more than 20 times as likely to demonstrate the proficiency in comparison to the HSE trained group and the difference was statistically significant (Ext (B) =20.25, 95% CI=3.91-105, p<0.000)."

>2 times as likely has been removed from the Figure 4

9. Kappas. Interrater reliability was 85%, but what was the kappa?

We used the traditional method of inter-rater reliability assessment, i.e., actual agreement between raters. The reviewer is correct that the kappa statistic is frequently used to test interrater reliability. The importance of rater reliability lies in the fact that it represents the extent to which the data collected in the study are correct representations of the variables measured. While there have been a variety of methods to measure interrater reliability, traditionally it was measured as percent agreement, calculated as the number of agreement scores divided by the total number of agreement + disagreements.(1, 2). In 1960, Jacob Cohen critiqued use of percent agreement due to its inability to account for chance agreement. He introduced the Cohen's kappa, developed to account for the possibility that raters actually guess on at least some variables due to uncertainty. Like most correlation statistics, the kappa can range from −1 to +1. While the kappa is one of the most commonly used statistics to test interrater reliability, it has limitations. One of these is the acceptable level of agreement. Furthermore, kappa is similar to a correlation coefficient and is not a direct

measure of agreement. For example a kappa = 0.85 indicates strong agreement between raters and interpreted by some readers as 85% agreement. This is incorrect. 0.85 squared (i.e., 0.722 = 72%), the percentage of variance explained is the closest to percentage agreement. Thus, kappa cannot be directly interpreted and it has become common for researchers to accept low kappa values in their interrater reliability studies. This is not acceptable in a healthcare context particularly in a high stakes assessment context (3, 4, 5).

The percent agreement statistic is easily calculated and directly interpretable. Its key limitation is that it does not take account of the possibility that raters guessed on scores. We controlled for the probability of guessing by assessors by training them on using the metrics in advance of scoring the study data. Raters were not permitted to assess study videos until they demonstrated an IRR > 0.8 consistently.

The kappa value of IRR = 0.85 would be 0.922. This however adds nothing to the understanding of how reliably the video recorded performances were assessed. The more important information is that raters were trained to use the metrics in advance, none of the assessments fell below an IRR < 0.8 the international accepted IRR level (6).

References

(1). Kazdin AE. Behavior modification in applied settings . Pacific Grove, CA: Brooks: Cole Publishing Co, 1994.

(2). Gallagher AG, O'Sullivan GC. Fundamentals of Surgical Simulation; Principles & Practices London: Springer Verlag 2011.

(3). Gallagher AG, O'Sullivan GC, Neary PC, et al. An objective evaluation of a multi-component, competitive, selection process for admitting surgeons into higher surgical training in a national setting. World J Surg 2014;38(2):296-304.

(4). Gallagher AG, Ritter EM, Satava RM. Fundamental principles of validation, and reliability: rigorous science for the assessment of surgical education and training. Surg Endosc 2003;17(10):1525-9. doi: 10.1007/s00464-003-0035-4 [published Online First: 2003/09/23]

(5). Gallagher AG, O'Sullivan GC, Leonard G, et al. OSATS and checklist scales reliability compared for high stakes assessments. The Australian and New Zealand Journal of Surgery 2014;84(7-8):568-73. doi: https://doi.org/10.1111/j.1445-2197.2012.06236.x [published Online First: 03 September 2012]

(6). Cooper C. Individual differences: Arnold London 2002.

10. p-values have been provided in table 1

Editorial

Redundant lines on page 16 removed

Reviewer 2

Thank you for your feedback

1.More in depth discussion added to both the introduction and discussion sections to justify the use of proficiency based progression as a pedagogical approach

2. Compared to standard training has been removed from the title

Reviewer 3

1. Yes the study is about addressing the effectiveness of PBP delivery as opposed to the components of NEWS/ISBAR this has now been more clearly outlined in the objectives section. More in depth discussion added to both the introduction and discussion sections to justify the use of proficiency based progression as a pedagogical approach

2. Reference format has been corrected

3. Power calculation-The power calculations were based on 1) peer reviewed, published clinical studies, 2) a scientific study on the transfer of training effect which showed a 42% effect on performance errors after simulation training and 3) the results from a previous pilot study using the same training and assessment methodology

i) The Power calculations are based on conservative estimates (i.e., >40%) from previous studies using the exact same proficiency based progression methodology. In these studies the effect size demonstrated a

- 74% difference (reference 1) Seymour et a
- 49% difference (reference 2) Cates, Lonn and Gallagher)
- 56% difference (reference 3) Angelo et al
- 54% difference (reference 4)Srinivasan et al

between the PBP trained group and the Control group

ii) The results of a previous study on the transfer of training effect showed a 42% effect on performance errors11 were used as a conservative guide to estimate effect size expected in the current study

iii) The results from a previous pilot study were extremely helpful in helping to guide effect size.

If anything, the effect size and power estimates were a strength of the study. The results of this study have also confirmed the robustness of the effect size (i.e., >40%) with PBP training. Furthermore, a recently published study by Srinivasan et al.10 has shown that this effect translates into clinical outcomes. The PBP trained grouped in the Srinivasan et al.,10 study showed a 53% reduction in epidural failure rates in comparison to the control group.

A sentence has ben added to the conclusion around numbers and effect size.

References

(1). Seymour NE, Gallagher AG, Roman SA, et al. Virtual reality training improves operating room performance: results of a randomized, double-blinded study. Ann Surg 2002;236(4):458-63; discussion 63-4. doi: 10.1097/01.SLA.0000028969.51489.B4 [published Online First: 2002/10/09]

(2). Cates CU, Lönn L, Gallagher AG. Prospective, randomised and blinded comparison of proficiency-based progression full-physics virtual reality simulator training versus invasive vascular experience for learning carotid artery angiography by very experienced operators. BMJ Simulation and Technology Enhanced Learning 2016:bmjstel-2015-000090.

(3). Angelo RL, Ryu RK, Pedowitz RA, et al. The Bankart performance metrics combined with a cadaveric shoulder create a precise and accurate assessment tool for measuring surgeon skill. Arthroscopy: The Journal of Arthroscopic & Related Surgery 2015;31(9):1655-70.

(4). Srinivasan KK, Gallagher A, O'Brien N, et al. Proficiency-based progression training: an 'end to end'model for decreasing error applied to achievement of effective epidural analgesia during labour: a randomised control study. BMJ open 2018;8(10):e020099.

Reviewer 4

Thank you for your feedback

1. Aims moved to the introductory section and more specific objectives added

2. Examples of a case and benchmark rubric given in tables 2 and 3

3. Greater detail added to the conclusion section and moved to discussion

Reviewer 5

Thank you for your feedback

1. The sentence on randomisation has been expanded

2. The training session lasted 3.5 hours in both E+S and E+PBP group this has been more clearly outlined for the E+PBP group in the methodology section the following added to part (iii) for that group -"The training session was 3.5 hours in duration, participants were required to stay until the end of the training regardless of progress. If an individual had completed all the cases, they were asked to assist by continuing to be the recipient of phone calls for their partner or by continuing to practice by repeating the cases if required"

3. P values added to demographics table 1

4 Limitations of the single centre design outlined in the discussion. Same 2 facilitators facilitated both the E+S and E+PBP groups although a larger study with multiple sites would undoubtedly make more robust.

3. Sentence added to discussion about single centre design and generalisability. Also how it has been used in the undergraduate curriculum.

4. The pilot study is being written up at present as it is a lower quality to study it was felt that this current study should be published first.

VERSION 2 – REVIEW

REVIEWER	Daryl Cheng Consultant Paediatrician / EMR Consultant The Royal Children's Hospital Melbourne, Australia
REVIEW RETURNED	05-Mar-2019

GENERAL COMMENTS	Introduction Page 7 Line 37 - this sentence has grammatical errors. Line 50 - shift-based patterns of work pg 8 Line 14 - I am not sure that there is a widespread desire to use communication tools. I think much of it has been mandated from a safety perspective to provide some structure around communication; and this needs to be partnered with education and training (as the authors have mentioned) Discussion Can the authors comment on interplay between various disciplines or craft groups? Difference between nursing and medical students? Do the authors think that there would be any difference if this study was performed in medical residents / nurses (as opposed to students?) Any future research implications? What about the impact on resources in terms of being able to deliver this type of proficiency based criteria?

REVIEWER	Alisa Khan, MD, MPH Boston Children's Hospital, USA
REVIEW RETURNED	16-Mar-2019

GENERAL COMMENTS	One minor comment is to consider moving the explanation of the nonsignficant E vs E+S logistic regression from the results to the discussion (or omitting entirely) as it is distracting, not necessary for the main conclusion around the effectiveness of the E+PBP program, and editorializes a bit more than is appropriate for a results section. Furthermore, with a p=.46, I don't believe it's accurate to say that it trended towards improvement.

REVIEWER	Dr. Zoe Hoare NWORTH CTU Bangor University
REVIEW RETURNED	17-Mar-2019

GENERAL COMMENTS	One minor comment might be to refer to the three groups design and multiplicity of testing within the power calculations and the subsequent analysis. I was not infering that p-values should be added to Table 1, in fact this could be considered to be pointless, see http://www.consort-statement.org/checklists/view/32--consort-2010/510-baseline-data. My reference to analysis including the demographics variables was to consider inclusion of any of these factors as possible covariates in such analysis not to test whether they were fundamentally different

	at baseline. Having re-considered the varibales collected though I do not think that this would necessarily have any value at this stage, First language and nationality are unlikely to be big factors in affecting the result and while gender may be the split in the recruited sample between male and female will probably not provide enough discrimination, likewise with age group. Personally I would remove the column of p-values for Table 1

VERSION 2 – AUTHOR RESPONSE

REVIEWER 3

Introduction

Page 7 Line 37 - this sentence has grammatical errors.

Line 50 - shift-based patterns of work

pg 8 Line 14 - I am not sure that there is a widespread desire to use communication tools. I think much of it has been mandated from a safety perspective to provide some structure around communication; and this needs to be partnered with education and training (as the authors have mentioned)

RESPONSE

The above comments refer to the original manuscript - these sections were already removed from the previous version The sentence now reads "Proficiency-based progression (PBP) training is a form of outcomes-based training that involves training individuals to achieve a proficiency benchmark. The process involves “deliberate” practice against a set of clearly defined objective metrics."

Discussion

Can the authors comment on interplay between various disciplines or craft groups? Difference between nursing and medical students? Do the authors think that there would be any difference if this study was performed in medical residents / nurses (as opposed to students?)

RESPONSE

The use of the undergraduate population is referred to in the discussion under weaknesses.

The relative impact on medical and nursing students is currently being written as a separate paper.

Any future research implications? What about the impact on resources in terms of being able to deliver this type of proficiency based criteria?

RESPONSE

Sentence modified in the discussion to read "There is a need for future research on the application of the programme in different clinical settings and its impact on patient outcomes"

Additional sentence added to the end of the discussion "Furthermore, improved performance with proficiency- based progression simulation was achieved with the same training time and facilitator/student ratio as standard simulation".

REVIEWER 1

Thank you for your revisions. The manuscript is much clearer now and the additional tables are extremely helpful.

One minor comment is to consider moving the explanation of the nonsignficant E vs E+S logistic regression from the results to the discussion (or omitting entirely) as it is distracting, not necessary for the main conclusion around the effectiveness of the E+PBP program, and editorializes a bit more than is appropriate for a results section. Furthermore, with a p=.46, I don't believe it's accurate to say that it trended towards improvement.

RESPONSE

This section removed as requested.

REVIEWER 5

One minor comment might be to refer to the three groups design and multiplicity of testing within the power calculations and the subsequent analysis.

RESPONSE

Multiple testing refers to any instance that involves the simultaneous testing of several hypotheses, e.g., in a repeated measures design. In our study only one primary hypothesis was tested (i.e., the ability to reach the proficiency benchmark on the standardised high-fidelity simulation assessment case). There were no repeated measure assessments. We used the same data for our analysis with no subset analysis for the testing of the main hypothesis. We therefore concluded that no correction is required for multiple testing.