Development of the TeamOBS-PPH – targeting clinical performance in postpartum hemorrhage

LISE BROGAARD¹, LONE HVIDMAN², KIM HINSHAW³, OLE KIERKEGAARD¹, TANJA MANSER⁴, PETER MUSAUE⁵, JULIE ARAFEH⁶, KAY I. DANIELS⁷, AMY E. JUDY⁷ & NIELS ULDBJERG²

¹Department of Obstetrics and Gynecology, Regional Hospital, Horsens, ²Department of Obstetrics and Gynecology, Aarhus University Hospital, Skejby, Denmark, ³Department of Obstetrics and Gynecology, Sunderland Royal Hospital, Sunderland, UK, ⁴Institute for Patient Safety, Faculty of Medicine, University Hospital Bonn, Bonn, Germany, ⁵Center for Health Sciences Education, INCUBA Science Park, Skejby, Denmark, ⁶Center for Advanced Pediatric and Perinatal Education, Stanford University, Palo Alto, CA, and ⁷Department of Obstetrics and Gynecology, School of Medicine, Stanford University, Palo Alto, CA, USA

Key words
Postpartum hemorrhage, health teams, interdisciplinary, validation studies, emergency teams, checklist, evaluation research, emergency care, hospital rapid response team

Correspondence
Lise Brogaard, Department of Obstetrics and Gynecology, Regional Hospital in Horsens, Sundvej 30, 8700 Horsens, Denmark. E-mail: brogaard@roed.eu

Conflict of interest
The authors have stated explicitly that there are no conflicts of interest in connection with this article.

Please cite this article as: Brogaard L, Hvidman L, Hinshaw K, Kierkegaard O, Manser T, Musaeus P, et al. Development of the TeamOBS-PPH – targeting clinical performance in postpartum hemorrhage. Acta Obstet Gynecol Scand 2018; 97:677–687.

Received: 26 September 2017
Accepted: 21 February 2018

DOI: 10.1111/aogs.13336

Abstract

Introduction. This study aimed to develop a valid and reliable TeamOBS-PPH tool for assessing clinical performance in the management of postpartum hemorrhage (PPH). The tool was evaluated using video-recordings of teams managing PPH in both real-life and simulated settings. Material and methods. A Delphi panel consisting of 12 obstetricians from the UK, Norway, Sweden, Iceland, and Denmark achieved consensus on (i) the elements to include in the assessment tool, (ii) the weighting of each element, and (iii) the final tool. The validity and reliability were evaluated according to Cook and Beckman. (Level 1) Four raters scored four video-recordings of in situ simulations of PPH. (Level 2) Two raters scored 85 video-recordings of real-life teams managing patients with PPH ≥1000 mL in two Danish hospitals. (Level 3) Two raters scored 15 video-recordings of in situ simulations of PPH from a US hospital. Results. The tool was designed with scores from 0 to 100. (Level 1) Teams of novices had a median score of 54 (95% CI 48–60), whereas experienced teams had a median score of 75 (95% CI 71–79; p < 0.001). (Level 2) The intra-rater (intra-class correlation (ICC) = 0.96) and inter-rater (ICC = 0.83) agreements for real-life PPH were strong. The tool was applicable in all cases: atony, retained placenta, and lacerations. (Level 3) The tool was easily adapted to in situ simulation settings in the USA (ICC = 0.86). Conclusion. The TeamOBS-PPH tool appears to be valid and reliable for assessing clinical performance in real-life and simulated settings. The tool will be shared as the free TeamOBS App.

Abbreviations: ICC, intra-class correlation; PPH, postpartum hemorrhage.

Introduction

Major postpartum hemorrhage (PPH), defined as blood loss of ≥1000 mL within 24 h of delivery, has an incidence of 4% and is the most frequent cause of maternal death worldwide (1,2). The risk of maternal death by PPH can be reduced significantly if the clinical team performs well (3–6). A clinical team’s management of PPH

Key Message

New clinical performance tool for assessing postpartum hemorrhage management. Use the free App TeamOBS to encourage structured feedback and learning of real-life postpartum hemorrhage.
The present study aimed to close this gap: addresses only an individual and not the performance of laceration. However, technical skills evaluation usually example, intrauterine palpation or suturing a cervical ing teaching and the training of technical skills, for their success. define high performing teams and discover the key to (14,17,18). Furthermore, in research, it may help us to feedback after simulation training or after a real-life PPH also be valuable in education, where it can be included in regions (9,13). Clinical performance assessment will a tool will be valuable for quality assessment and bench- a tool that also assesses clinical performance. Using such tools, we need a tool that also assesses clinical performance. Using such a tool will be valuable for quality assessment and benchmarking of performance over time and between hospitals or regions (9,13). Clinical performance assessment will also be valuable in education, where it can be included in feedback after simulation training or after a real-life PPH (14,17,18). Furthermore, in research, it may help us to define high performing teams and discover the key to their success. The present study aimed to close this gap: • by developing a tool for assessing the clinical performance of teams managing PPHs in both real-life situations and in situ simulations.

Material and methods
Development of the TeamOBS-PPH tool
In the first part of this study, we conducted an e-mail-based Delphi process to develop the TeamOBS-PPH tool (Figure 1). The Delphi panel consisted of 12 senior obstetricians from maternity units in the UK (n = 3), Norway (n = 2), Sweden (n = 3), Denmark (n = 3), and Iceland (n = 1). In four rounds, the experts answered questionnaires (Supporting Information Appendix S1) concerning the items to be included in the tool and the weight to be assigned to each item (Figure 2). Based on these results (19,20), we developed the TeamOBS-PPH tool for assessment of clinical performance (Figure 3). The tool generates a score from 0 to 100 with a minimal pass level of 60, below which the risk of harming the patient is to be considered. We also defined a lower level of high performance at 85, above which no or only minor errors affected the score.

Testing of validity and reliability
In the second part of this study, to evaluate the validity and reliability, we used the framework and conceptual definitions devised by Cook and Beckman (2006), described as the five sources of validity evidence (21,22). To answer our validation questions as listed in Table 1, tests were performed at three levels.

Level 1: Performance in test setting. The initial testing of TeamOBS-PPH was conducted from November 2014 to February 2015. We used four selected video-recordings of in situ team trainings of simulated severe PPH (1300 mL) due to refractory uterine atony from two Danish hospitals: the Regional Hospital in Horsens – Maternal Care Level 2 (HEH), and Aarhus University Hospital – Maternal Care Level 3 (AUH). The two Expert teams consisted of five to six clinicians: obstetricians, midwives, and technicians. The two Novice teams consisted of midwifery and medical students in their final year of training. The simulations were performed on the delivery unit with a hybrid mannequin (PROMPT-Birth- ing Simulator, Limbs and Things).

We recruited four obstetric providers as raters (one registrar, one senior consultant, one junior midwife, and one senior midwife). These providers were not familiar with the tool as they were not members of the Delphi panel. They were formally trained as “raters” in a one-
Based on literature the TeamOBS group identified the important items.

**Result**
A list with 19 items

The expert group consisted of 12 experts from different departments from five countries.

**Iceland**
1: Reynir T. Geirsson, Professor, MD, PhD, FRCOG, University of Iceland, Reykjavik

**Norway**
2: Birgitte Sanda, Consultant Obstetrician (CO). Sarlandet Hospital Trust, Kristiansand
3: Torbjoern Moe Eggeboe, Professor, PhD, CO. St. Olav’s Hospital, Trondheim

**Sweden**
4: Charlotta Grunewald, Associate Professor, CO. Director of Obstetric unit, Karolinska University Hospital, Stockholm
5: Jan Leyon, CO. Head of delivery department, Skaraborg Hospital, Skövde

**Denmark**
6: Nana Wiberg, PhD, CO. University Hospital Lund
7: Lise Lotte T. Andersen, CO. OUH, Odense
8: Lone Krebs, CO. DMSc, Associate professor, Holbaek Hospital
9: Jette Led Sorensen, PhD, MMed, Associate professor, CO. Rigshospitalet, Copenhagen

**United Kingdom**
10: Kim Hinshaw, FRCOG, CO. Visiting Professor, Sunderland, Royal Hospital, Tyne & Wear
11: Tim Draycott, FRCOG, CO. North Bristol NHS trust
12: Philip J. Steer, FRCOG, Emeritus Professor, Imperial College London Chelsea and Westminster Hospital

The expert group was asked to add, remove or suggest corrections to the 19 items.

**Result**
4 items adjusted; none added or removed.

Weights of elements. Each member of the expert group weighted the importance of each item on a 5-point Lickert scale.

**Result**
Consensus was achieved on 17 of the 19 items.

The expert group weighted the same items again, this time with an anonymous summary of the scores from Delphi round 2.

**Result**
Experts achieved consensus on 19 of the 19 items.

After preliminary testing the final assessment tool TeamOBS-PPH was approved.

**Figure 1.** Delphi process. [Color figure can be viewed at wileyonlinelibrary.com].
The TeamOBS-PPH tool

Intravenous fluid

Delphi round 2

1. Not important 0%
2. Less relevant 0%
3. relevant 0%
4. Important 16.7%
5. Essential 83.3%

Delphi round 3

1. Not important 0%
2. Less relevant 0%
3. relevant 0%
4. Important 0%
5. Essential 100%

Tranexamic acid

Delphi round 2

1. Not important 8.3%
2. Less relevant 25%
3. relevant 50%
4. Important 16.7%
5. Essential 0%

Delphi round 3

1. Not important 0%
2. Less relevant 16.7%
3. relevant 25%
4. Important 58.3%
5. Essential 0%

Figure 2. Consensus of opinion. Examples of changes in agreement within the Delphi panel from round two to three. Concerning intravenous fluid, we obtained 100% agreement. Concerning tranexamic acid, we achieved agreement defined as at least 90% of the experts’ scores falling within three neighboring categories on the Likert scale. [Color figure can be viewed at wileyonlinelibrary.com].

During training, raters also jointly assessed a video-recording of a team from an in situ simulation. After training, the four raters applied the TeamOBS-PPH tool to assess all four video-recordings independently. After 1 month, the four raters reassessed all video-recordings.

Level 2: Real-life setting, Denmark. We obtained video-recordings of teams managing real-life PPH situations at two Danish hospitals: HEH and AUH. In all 17 delivery rooms, two or three mini-dome high-definition surveillance cameras were placed in the corner at ceiling level to capture a view of the clinical team from the head-end facing the patient to minimize inappropriate patient exposure. However, staff and patient faces were potentially identifiable. A microphone was placed at the center of the ceiling. In each delivery room, we installed a video-recording system which was activated by a Bluetooth chip in the obstetrician’s telephone when he or she entered the room. Camera capacity allowed recording day and night with a storage capacity of 5 min. When the system was activated upon the obstetrician’s entry, the immediately preceding 5 min of video-recording were saved to the server, and all subsequent footage was captured. The video was deleted after 48 h.

The midwives reported eligible cases (PPH ≥1000 mL) to the research team. The midwives’ reports included the total blood loss (mL) assessed by weight. Thus, the research team had 48 h to obtain consent from all parties to download the video-recording from the server for research purposes. Inclusion of teams is visualized in Figure 4.

We recruited two senior consultant obstetricians from the studied hospitals (HEH and AUH), who had not participated in the Delphi process or in the Level 1 testing. They had the same formal one-hour training as the raters performing the Level 1 testing, and they subsequently assessed all 85 video-recordings independently, blinded to
each other's scores and the reported total blood loss. After every five recordings, the raters were asked to discuss any difficulties they had experienced but they were not allowed to alter any prior assessments. After 1 month, the raters reassessed 20% of the recordings (randomly selected) to evaluate intra-rater agreement over time. Videos were collected during 15 months in 2014–2015. The videos were analyzed in the spring of 2016.

**Level 3: External validity, USA.** The TeamOBS-PPH tool is based on traditional approaches in five countries in northern Europe. To explore whether our results can be applied in a different cultural context, we tested its use at the Obstetric Center at Lucile Packard Children’s Hospital, Stanford, California, USA. Simulation was used because it was not possible to find another hospital conducting live video-recordings of PPH that could be used for research.

To account for system differences, a simplified Delphi process was conducted as a group discussion among five clinicians working in the obstetric field (one Fellow, two Consultant Obstetricians, one Consultant Obstetric Anesthetist, and one Labor and Delivery Nurse). They were asked for approval or corrections to each item considering standard clinical practice in their regional area. The final tool was adjusted and approved by the Delphi group.

Video-recordings of simulated in situ team training of how to manage severe PPH (1500 mL) due to refractory uterine atony were collected in the same hospital in 2011–2016. Each recording was of teams consisting of 7–10 clinicians: obstetricians, anesthesiologist, labor and delivery nurses, and technicians. Fifteen video-recordings were included. Six of the teams were performing on the labor and delivery unit using a hybrid mannequin (Mama Birthing Simulator, Laerdal). Nine teams performed the simulation in the operating theater with the mannequin NOELLE (Maternal and Neonatal Birthing Simulator from Gaumard).

Two raters (one Fellow and one Consultant Obstetrician) assessed the 15 video-recordings independently and blinded to each other’s scores. They underwent the same formal training as all other raters. Analysis was conducted in the spring of 2016.
Statistical analysis

The statistical analysis of the clinical performance scores was performed on the logit-transformed scale using the normal model and back-transformed using the inverse logit function (24,25). Further description of Bland–Altman plots can be found in Supporting Information Figure S1. Analysis of blood loss was performed using the normal model after log-transformation. The relation between clinical performance and total blood loss was analyzed using simple linear regression. The potential confounding of bleeding velocity (mL/min) was assessed using multiple linear regression analysis. The model was checked by diagnostic plots of residuals. An intra-class correlation (ICC) >0.75 was considered high agreement (26). We used STATA version 14.0 for the statistical analysis (StataCorp, College Station, TX, USA).

Table 1. Argument for validity.

| Five sources of validity | Validity question? | Data | Method | Results |
|-------------------------|--------------------|------|--------|---------|
| Content evidence        | Measures what it was intended to? | Delphi process | Consensus of items, weight of importance, and final tool TeamOBS-PPH TeamOBS-PPH used a rating scale with five categories, weighed items, and a global rating scale. After 1 h of training, the midwives and obstetricians were comfortable with the tool. The App solution facilitates ease of assessment. |
| Response process evidence | Easy to use and understand? | Rater handbook | Systematic feedback |
| Internal structure evidence | Distinguishes high and low performance? | Teams of either novices or experienced | Compare scores | Novices, mean S4 (95% CI 48–60) Experienced, mean 75 (95% CI 71–79), p ≤ 0.001 1 rater, ICC = 0.72 (95% CI 0.59–0.80) Average of 2 raters, ICC = 0.83 (95% CI 0.75–0.89) |
| Reproducibility? | Inter-rater agreement | Two raters scored 85 videos of real-life teams, Denmark | ICC = 0.72 (95% CI 0.59–0.80) |
| | Intra-rater agreement | Reevaluation >1 month | 1 rater, ICC = 0.93 (95% CI 0.86–0.96) Average of 2 raters, ICC = 0.96 (95% CI 0.92–0.98) |
| | Across the items in TeamOBS-PPH | ICC recalculated after each item had been deleted | Only one item, "Consider the cause of bleeding", changed the ICC (increased by only 0.02) 1st half, ICC = 0.80 (95% CI 0.62–0.89) 2nd half, ICC = 0.86 (95% CI 0.73–0.92), p > 0.05 |
| | Rater drift | Compare first 43 videos with last 42 videos. | Across scenarios Causes of bleeding | Tool was applicable in all cases: uterine atony, retained placenta, or lacerations |
| | | | Individual ICC = 0.75 (95% CI 0.41–0.91) Average ICC = 0.86 (95% CI 0.58–0.95) |
| 4. Relation to other variable evidence | Is the performance associated with blood loss? | Amount of blood loss in total | Total amount of bleeding in the cases | Clinical performance was significantly associated with the amount of blood loss (Table 2) |
| | | | | Below 50: 0 cases = 0% Below 60: 3 cases = 2% Below 70: 5 cases = 8% |
| 5. Consequence evidence | The team fails | Different levels of passing score | Total of 85 videos used above for inter-rater agreement | Eight cases (10%) consensus discussion |
| | | | | |

ICC, intra-class correlation.
Ethical approval

All participants gave written consent for the videos to be analyzed for research. Separate approval was necessary for video-recording of real-life teams. The study was approved in May 2014 by the legal department of the Central Denmark Region, the Danish Data Protection Agency (2012-58-006), and the Research Foundation of The Central Denmark Region (case no. 1-16-02-257-14). Further information about ethical and legal requirements is contained in Supporting Information Appendix S2.

Results

Part 1: The TeamOBS-PPH tool development

All 12 experts completed the Delphi process. In the first round, they adjusted four items but did not add or remove any items. One example of adjustment was calling for delivery of blood, which was adjusted to include both whole blood and packed cells to allow for differences between local guidelines in various countries. In the second round, the experts reached consensus on 17 of the 19 items; in the third round, on all 19 items. The Delphi panel approved the final tool as shown in Figure 3.

Part 2: Validity and reliability

A complete list of validity process results is presented in Table 1.

Level 1: Performance in test setting. The novice teams obtained a median score of 54 (95% CI 48–60) and the expert teams a median score of 75 (95% CI 71–79); \( p \leq 0.001 \). The four raters assessed four videos from Level 1 and four videos from Level 2; the inter-rater agreement of one rater was ICC = 0.83 (95% CI 0.59–0.96); the average of four raters was ICC = 0.95 (95% CI...
re-evaluation after 1 month of one rater was ICC = 0.88 (95% CI 0.71–0.96), the average of four raters was ICC = 0.94 (95% CI 0.83–0.97). Based on the excellent inter-rater agreement, we decided to use two raters in Level 2.

**Level 2: Real life setting, Denmark.** We included 85 of the 188 cases of severe PPH occurring among the 6700 deliveries performed at the two hospitals in 2015. Reasons for non-eligibility are given in Figure 4. The average amount of bleeding among the included cases was a median of 1648 mL (95% CI 1542–1760) compared with a median of 1641 mL (95% CI 1524–1767), p = 0.93 among the excluded cases. Eighty-one different team combinations were included, four of which appeared in two recordings. The average team size was five (physicians, midwives, and technicians).

Rater 1 scored with a median of 87 (95% CI 85–89), range 45–100; rater 2 with a median of 89 (95% CI 87–91), range 59–99 (p = 0.99). The inter-rater agreement of two raters and 85 videos was 0.83 (95% CI 0.74–0.89). The intra-rater agreement after 1 month for two raters was ICC = 0.96 (95% CI 0.92–0.98). The correlation between the weighted scores and the patient safety scores is visualized in Figure 5a. The difference in agreement is visualized with limits of agreement in Figure 5b. The tool was applicable in all 85 cases of PPH, irrespective of the cause (uterine atony, retained placenta or lacerations).

Only one item, 3.1 “consider the cause of the bleeding” changed the consistency when re-calculated after each item had been deleted (27). Deleting this item increased the ICC by 0.02 to a total ICC of 0.85, meaning that this variable is reliable.

The total blood loss in the 85 real-life teams was associated with the level of performance: A score of 60 was associated with a median blood loss of 2097 mL (95% CI 1696–2593 mL), a score of 85 with 1696 mL (95% CI 1577–1824 mL), and a score of 100 with 1493 mL (95% CI 1315–1695 mL). The difference between the three levels of performance was significant (p = 0.0029) (Table 2). Our results remained stable when adjusting for bleeding velocity as a potential confounder (Supporting Information Table S1).

**Level 3: Simulation setting, USA.** In the United States, the Delphi panel adjusted the phrasing of two items: “Blood test for FBC, blood type for compatibility and cross match” was rephrased as “Blood test for CBC, coags, cross match and type”, and “Monitor observations: pulse, blood pressure, and respiratory rate” was replaced with “Monitor observations: pulse, blood pressure, pulse ox, and respiratory rate”. One item, “Documentation on PPH chart”, was adjusted in weight from 2.5 to 3.5. A new item was added: “Correct placement of the Bakri balloon”. No items were deleted. Following these adjustments, the Delphi panel approved the tool for local/
regional use. Rater 1 scored with a median of 79.3 (95% CI 71.8–85.4), range 47–92, and rater 2 with a median of 79.1 (95% CI 72.4–84.6), range 58–93; \( p = 0.92 \). The inter-rater ICC was 0.86 (95% CI 0.58–0.95; two raters, 15 recordings).

**Usability of the TeamOBS-PPH tool**

The tool was easy to use. At all three levels, the raters were trained for one hour and then used the tool as the video was rolling, i.e. they did not see the video more than once.

**Discussion**

This study is the first study to develop a tool for assessment of the clinical performance of teams managing PPH. The TeamOBS-PPH tool includes an objective weighted score based on a checklist including 19 items and a more subjective overall patient safety score. The tool is applicable in both real-life and simulated settings. The reliability and validity are high, and the tool is adaptable and modifiable to local clinical guidelines.

The main strength of this study is that it is based on video-recordings of the clinical management of patients with real-life PPH. To our knowledge, this has not been done previously. Another strength is that the tool was developed through a Delphi process with a panel of experts from five countries, and we were able to show that the tool was valid cross-culturally. Furthermore, higher clinical performance was significantly associated with less bleeding.

A limitation is that, despite the high technical quality of the video-recordings, it was difficult to accurately assess the amount of blood loss, which is why the raters had to rely on the team itself to verbalize this variable.

The risk of response bias must be considered, as we deliberately chose raters from the study population in order to utilize their knowledge of the actual health organization. Furthermore, we included the more subjective global score, contributing not less than 50% of the total score (14,28). The reason for this decision was that the global score quite often further decreases lower weighted scores and increases higher weighted scores (Figure 5a).

Based on our findings, we recommend that:

- Each PPH recording be assessed by two independent raters. Two is sufficient as the median difference between the observers was only 3 points, which is far below our predefined acceptable difference of 15 points for a single event (Figure 5b). By use of less experienced raters, the inter-rater agreement might be lower; therefore, three raters could be considered.
- If the difference between the two raters exceeds 15 points, the two raters should compare their ratings. This happened in our study in eight of 85 cases (Figure 5b), primarily among scores below 80. Most often the raters realized that one of them had misinterpreted an item.
- The acceptable level of performance be set at 60. To ensure a fair judgement of the few cases with lower scores, we the raters should also discuss and agree on these ratings.
- The tool be modified according to local clinical guidelines. For example, the US teams checked the uterine cavity and considered placement of an intrauterine balloon in the delivery room, whereas in Northern Europe this treatment would generally take place in the operating theater.
- The raters consider using the free TeamOBS-PPH App, which may help with assessment.
- The TeamOBS-PPH tool be used for post-event assessment, i.e. summative evaluation of performance. It is not designed to be used as an alternative for a procedural checklists or as a cognitive aid for teams to follow during the event (16).

We suggest that the TeamOBS-PPH be considered for use in the following situations:

- Education, as a feedback tool after simulation training to encourage reflective practice. The App version of the TeamOBS-PPH tool includes a feedback module, in which the performance in different categories is visualized in graphs.
- Structured debriefing after real-life clinical events. We suggest using the feedback module in the App, where the model “SHARP - 5-step feedback and debriefing tool (29)” is included.
- Quality assurance. This step is important to facilitate the best possible conditions in the delivery wards, allowing teams to become more effective and improve patient safety.
- Research (7), for example, to analyze the association between clinical performance (performing tasks) and

---

**Table 2.** Clinical performance and blood loss.

| Level of performance | TeamOBS-PPH score | Total blood loss, median 95% CI |
|----------------------|-------------------|--------------------------------|
| Fail                 | <60               | NA                             |
| Minimal pass         | 60                | 2,097 mL 1696–2593 mL          |
| High performance     | 85                | 1,696 mL 1577–1824 mL          |
| Excellent            | 100               | 1,493 mL 1315–1695 mL          |

These figures are based on the 85 recordings from real-life settings. NA, not available due to low number of cases. Difference in blood loss between the three groups, \( p = 0.029 \).
non-technical performance (communication, leadership, decision-making, etc.) (12).

In conclusion, our study provides a new tool for assessing clinical performance in the management of PPH. It was developed through an international Delphi process and tested in real-life PPH with acceptable validity. The TeamOBS-PPH tool offers an opportunity to assess actual team performance which is useful for education, research, and feedback. We hope that the App TeamOBS will serve as an aid for the busy clinician to ease the process of providing structured feedback and to encourage continuous learning to improve team performance during real-life PPH.

Acknowledgments
We are grateful to all participants in the Delphi panel: Reynir T. Geirsson, Birgitte Sanda, Torbjørn Moe Eggeboe, Charlotta Grunewald, Jan Leyon, Nana Wiberg, Lise Lotte T. Andersen, Lone Krebs, Jette Led Sorensen, Kim Hinshaw, Tim Draycott, and Philip J. Steer. We are also grateful for the statistical assistance by Professor Erik Thorlund Parner, Department of Biostatistics, Aarhus University, Denmark.

Funding
We thank following organizations and departments for funding: Tryg Foundation (Trygfonden) (grant ID no. 109507); the Regional Hospital in Horsens, Department of Obstetrics and Gynecology; the Regional Postgraduate Medical Education Administration Office North.

References
1. Knight M, Callaghan WM, Berg C, Alexander S, Bouvier-Colle M-H, Ford JB, et al. Trends in postpartum hemorrhage in high resource countries: a review and recommendations from the International Postpartum Hemorrhage Collaborative Group. BMC Pregnancy Childbirth. 2009;9:55.
2. Carroli G, Cuesta C, Abalos E, Gulmezoglu AM. Epidemiology of postpartum haemorrhage: a systematic review. Best Pract Res Clin Obstet Gynaecol. 2008;22:999–1012.
3. Clark SL. Strategies for reducing maternal mortality. Semin Perinatol. 2012;36:110–20.
4. Leduc D, Senikas V, Lalone A, Ballerman C, Biringer A, Delaney M, et al. Active management of the third stage of labour: prevention and treatment of postpartum hemorrhage. J Obstet Gynaecol Can. 2009;31:980–93.
5. Mousa HA, Blum J, Abou El Senoun G, Shakur H, Alfirevic Z. Treatment for primary postpartum haemorrhage. Cochrane Database Syst Rev 2014;(2):CD003249.
6. Joint Commission on Accreditation of Healthcare Organizations, USA. Preventing maternal death. Sentinel Event Alert. 2010;44:45–7.
7. Manser T. Team performance assessment in healthcare: facing the challenge. Simul Healthc. 2008;3:1–3.
8. Siassakos D, Fox R, Bristowe K, Angourl J, Hambly H, Robson L, et al. What makes maternity teams effective and safe? Lessons from a series of research on teamwork, leadership and team training. Acta Obstet Gynecol Scand. 2013;92:1239–43.
9. Wiss V, Van Der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. Lancet. 2001;357:945–9.
10. Kaye AD, Okanlawon OJ, Urman RD. Clinical performance feedback and quality improvement opportunities for perioperative physicians. Adv Med Educ Pract. 2014;5:115–23.
11. Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C, et al. Objective structured assessment of technical skills (OSATS) for surgical residents. Br J Surg. 1997;84:273–8.
12. Morgan PJ, Tregunno D, Pittini R, Tarshis J, Regehr G, Desousa S, et al. Determination of the psychometric properties of a behavioural marking system for obstetrical team training using high-fidelity simulation. BMJ Qual Saf. 2012;21:78–82.
13. Goddard M, Davies HT, Dawson D, Mannion R, McInnes F. Clinical performance measurement: part 1 – getting the best out of it. J R Soc Med. 2002;95:508–10.
14. Schmutz J, Manser T, Keil J, Heimberg E, Hoffmann F. Structured performance assessment in three pediatric emergency scenarios: a validation study. J Pediatr. 2015;166:498–504.e1.
15. Bajaj K, Rivera-Chiauzzi EY, Lee C, Shepard C, Bernstein PS, Moore-Murray T, et al. Validating obstetric emergency checklists using simulation: a randomized controlled trial. Am J Perinatol. 2016;33:1182–90.
16. Hilton G, Daniels K, Goldhaber-Fiebert SN, Lipman S, Carvalho B, Butwick A. Checklists and multidisciplinary team performance during simulated obstetric hemorrhage. Int J Obstet Anesth. 2015;25:9–16.
17. Isenbeng SB, McGaghie WC, Petrusa ER, Lee Gordon D, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. Med Teach. 2005;27:10–28.
18. Meriën AE, van de Ven J, Mol BW, Houterman S, Oei SG. Multidisciplinary team training in a simulation setting for acute obstetric emergencies: a systematic review. Obstet Gynecol. 2010;115:1021–31.
19. Morgan PJ, Lam-McCulloch J, Herold-McIlroy J, Tarshis J. Simulation performance checklist generation using the Delphi technique. Can J Anaesth. 2007;54:992–7.
20. Schmutz J, Eppich WJ, Hoffmann F, Heimberg E, Manser T. Five steps to develop checklists for evaluating clinical
performance: an integrative approach. Acad Med. 2014;89:996–1005.

21. Cook DA, Zendejas B, Hamstra SJ, Hatala R, Brydges R. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. Adv Health Sci Educ Theory Pract. 2014;19:233–50.

22. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane’s framework. Med Educ. 2015;49:560–75.

23. Menard MK, Kilpatrick S, Saade G, Hollier LM, Joseph GF, Barfield W, et al. ACOG/SMFM obstetric care consensus: levels of maternal care. Am J Obstet Gynecol. 2015;125:502–15.

24. Bland JM, Altman DG. Applying the right statistics: analyses of measurement studies. Ultrasound Obstet Gynecol. 2003;22:85–93.

25. Carstensen B. Comparing clinical measurement methods: a practical guide. Chichester: John Wiley & Sons, Ltd, 2010.

26. Khan KS, Chien PF. Evaluation of a clinical test. I: Assessment of reliability. BJOG. 2001;108:562–7.

27. Morgan PJ, Cleave-Hogg D, DeSousa S, Tarshis J. High-fidelity patient simulation: validation of performance checklists. Br J Anaesth. 2004;92:388–92.

28. Winckel CP, Reznick RK, Cohen R, Taylor B. Reliability and construct validity of a structured technical skills assessment form. Am J Surg. 1994;167:423–7.

29. Ahmed M, Arora S, Russ S, Darzi A, Vincent C, Sevdalis N. Operation debrief: a SHARP improvement in performance feedback in the operating room. Ann Surg. 2013;258:958–63.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Appendix S1. The Delphi process and development of the tool.

Appendix S2. Resumé of ethical and legal requirement.

Figure S1. Bland–Altman plots. The assumption of uniform mean, standard deviation (SD), and normality was not met on the original score: “clinical performance score”. The differences were not constant throughout the range of measurements because the scores were in percentages and variance was diminishing towards 100%. The assumption of uniform mean, SD, and normality was, however, satisfied on the logit-transformed scores. Figures visualize the Bland–Altman plots for the original scale (%) and the logit scale (logit). Since the scores are given in percentages, the assumptions for Bland–Altman and limit of agreement are not met. A solution would be to transform to logit function. As the logit model assumes that no differences are available for 100%, all data were reduced by 3% during the transformation. The transformed scale is not easily interpreted in a clinical sense Therefore, it is transformed back to the original scale where the limits of agreement are, of course, not straight lines. The result is Figure 5b.

Table S1. Clinical performance and total blood loss adjusted for bleeding velocity (mL/min).