European Journal of Obstetrics & Gynecology and Reproductive Biology 216 (2017) 184–191

Contents lists available at ScienceDirect

European Journal of Obstetrics & Gynecology and Reproductive Biology

journal homepage: www.elsevier.com/locate/ejogrb

Review article

Assessing teamwork performance in obstetrics: A systematic search and review of validated tools

Annemarie F. Fransen[a,b,*], Liza de Boer[c], Dieneke Kienhorst[d], Sophie E. Truijens[a], Pieter J. van Runnard Heimel[a], S. Guid Oei[a,e]

[a] Dept. of Obstetrics and Gynecology, Máxima Medical Centre, Eindhoven-Veldhoven, the Netherlands
[b] Dept. of Obstetrics and Gynecology, Maastricht University Medical Centre, Maastricht, the Netherlands
[c] Dept. of Obstetrics and Gynecology, Ikazia Hospital, Rotterdam, the Netherlands
[d] Dept. of Obstetrics and Gynecology, Academic Hospital Paramaribo, Suriname
[e] Dept. of Electrical Engineering, Eindhoven University of Technology, Eindhoven, the Netherlands

ARTICLE INFO

Article history:
Received 2 December 2016
Received in revised form 15 June 2017
Accepted 23 June 2017

Keywords:
Teamwork
Team performance
Obstetrics
Assessment tools
Simulation
Deliberate practice
Validation
Non-technical skills

ABSTRACT

Teamwork performance is an essential component for the clinical efficiency of multi-professional teams in obstetric care. As patient safety is related to teamwork performance, it has become an important learning goal in simulation-based education. In order to improve teamwork performance, reliable assessment tools are required. These can be used to provide feedback during training courses, or to compare learning effects between different types of training courses. The aim of the current study is to (1) identify the available assessment tools to evaluate obstetric teamwork performance in a simulated environment, and (2) evaluate their psychometric properties in order to identify the most valuable tool(s) to use. We performed a systematic search in PubMed, MEDLINE, and EMBASE to identify articles describing assessment tools for the evaluation of obstetric teamwork performance in a simulated environment. In order to evaluate the quality of the identified assessment tools the standards and grading rules have been applied as recommended by the Accreditation Council for Graduate Medical Education (ACGME) Committee on Educational Outcomes. The included studies were also assessed according to the Oxford Centre for Evidence Based Medicine (OCEBM) levels of evidence. This search resulted in the inclusion of five articles describing the following six tools: Clinical Teamwork Scale, Human Factors Rating Scale, Global Rating Scale, Assessment of Obstetric Team Performance, Global Assessment of Obstetric Team Performance, and the Teamwork Measurement Tool. Based on the ACGME guidelines we assigned a Class 3, level C of evidence, to all tools. Regarding the OCEBM levels of evidence, a level 3b was assigned to two studies and a level 4 to four studies. The Clinical Teamwork Scale demonstrated the most comprehensive validation, and the Teamwork Measurement Tool demonstrated promising results, however it is recommended to further investigate its reliability.

Contents

* Corresponding author at: Postbus 7777, 5500MB Veldhoven, the Netherlands.
E-mail address: annemariefransen@hotmail.com (A.F. Fransen).

http://dx.doi.org/10.1016/j.ejogrb.2017.06.034
0301-2115/

Introduction

There is more to clinical performance of obstetric teams than individual knowledge and technical skills. An essential component underlying the variation in clinical performance is teamwork [1]. In the current study we refer to teamwork when using the term "teamwork performance". This term stands for the performance of the teams as a collective, instead of the individual performance of the team members. It consists of teamwork behaviours, including interpersonal (e.g. communication, leadership) and cognitive skills (e.g. decision making, planning, situational awareness) [2]. Since teamwork performance in healthcare is associated with both effective and safe healthcare [3–5], education and research on this topic is needed.

A successful learning method to improve teamwork performance is simulation-based team training [6–8]. However, for such training to be effective, it should include reliable measurement of performance as described by the theory of deliberate practice [9,10]. To meet this requirement, validated assessment tools can be used. These tools support two purposes; on the one hand they provide objective feedback, on the other they enable reliable comparison between different types of team training courses. It is preferable that these tools are developed and validated within the medical specialty of interest. This is especially pertinent to obstetrics, where discipline-specific teamwork behaviours have been identified [11].

When using assessment tools, one should consider their psychometric characteristics. The Accreditation Council for Graduate Medical Education (ACGME) Committee on Educational Outcomes has proposed a set of standards, grading rules and summary rules for evaluating the quality of assessment methods [12]. The ACGME guidelines include standards of six topics: reliability, validity, ease of use, resources required, ease of interpretation, and educational impact. The desired weight of each topic depends on the specific assessment conditions [13].

The aim of the current study is to (1) identify available assessment tools to evaluate obstetric teamwork performance in simulated settings and (2) report on their psychometric properties in order to identify the most valuable tool(s) to use. The provided overview can support users while choosing an appropriate assessment tool, based on their educational needs. An additional advantage is reserved for tool developers, as this overview makes it possible to put future assessment tools into perspective.

Methods

Search

A systematic search was performed to select all available validated assessment tools for the evaluation of teamwork performance (at team level) in simulated settings. The search was performed with the assistance of a professional medical research librarian. The electronic databases of PubMed, MEDLINE and EMBASE were searched for articles between June 1975 and June 2016, without any language restrictions. Additional articles were identified by handsearching the references of included articles. The search strategy was included four categories: simulation, teamwork skills, assessment and obstetrics. For this search, also Medical Subject Headings (MeSH) were used. The complete search strategy can be found in Appendix A in Supplementary material.

Definitions

Teamwork performance is the performance measured at the level of the team, which concerns teamwork behaviours. These skills can be broadly divided into two groups: a) cognitive or mental skills (such as decision making, planning, and situational awareness) and b) social or interpersonal skills (such as teamwork, communication, and leadership) [14].

Simulated environments are medical simulation-based environments that resemble reality. They can be used for educational or qualification purposes, in which medical issues can be managed without the risks of real patient care.

In- and exclusion of studies

Titles and abstracts were screened to judge their eligibility for inclusion. Studies referring to assessment tools for teamwork performance in simulated settings were included. The validation process had to be presented. We excluded studies describing assessment tools for medical technical skills, teamwork performance in other medical specialties than obstetrics, and/or teamwork skills at the individual level. Likewise, tools validated for non-medical or undergraduate healthcare professionals were excluded. Conference abstracts and meeting proceedings were also excluded.

Selection of studies

Two reviewers independently reviewed the titles and abstracts. Full articles and references were retrieved whenever the reviewers could not decide on eligibility or when reviewers disagreed. A standardised coding form was created to indicate the reason for in- or exclusion and to collect handsearched references of interest. In case of any disagreement between reviewers, a third reviewer was consulted. In case of missing data or doubts about eligibility, authors were contacted by e-mail.

Data extraction

Data regarding tool characteristics, development, and validity of the tool were extracted. Psychometric properties, including validity and reliability measurements, were extracted with a standardised data form. Validity refers to the extent to which the tool is actually measuring what it claims to be measuring (e.g. construct validity, content validity). Reliability refers to the ability of the assessment tool to reproduce results under a same condition (e.g. inter-rater reliability, test-retest reliability). In addition to these psychometric properties, data concerning usefulness properties were collected. These properties are listed under the heading "Usefulness" and include: ease of use, resources required, ease of interpretation and educational impact (Appendix B in Supplementary material). The outcome measures of interest were retrieved from the full articles by two independent reviewers, in duplicate. Disagreement between reviewers was solved by consensus, and if necessary, by consulting a third reviewer.

Quality evaluation of included assessment tools

The guidelines proposed by the ACGME Committee on Educational Outcomes were applied to evaluate the assessment tools' quality. We have chosen to use these guidelines as they provided a set of standards for psychometric and usefulness properties (Appendix C in Supplementary material). Besides, they describe useful grading rules for an overall evaluation of the quality of the assessment method and the level of evidence (Table 1) [12]. Swing et al. mentioned that the ACGME guidelines are therefore more objective than other evaluation criteria [12]. According to the ACGME guidelines, the psychometric and usefulness properties possess the following topics: validity, reliability, ease of use, resources required, ease of interpretation, and educational impact. In the current study, two independent researchers applied the ACGME guidelines, using a standardised form. Subsequently, a grading for an overall recommendation and a level of evidence (Table 1) was assigned by each independent reviewer. Disagreement between reviewers was solved by consultation of the third reviewer. The two independent reviewers additionally assessed all included studies according to the Oxford Centre for Evidence Based Medicine (OCEBM) levels of evidence for diagnostic studies [15]. Disagreement between reviewers was solved by consensus.

Results

Search results

A total of 439 studies have been identified with the search (Fig. 1). Exploration of all references used in these reports yielded another 17 studies. After removing all duplicates 382 studies remained. Among these studies, 267 studies were excluded based on their title and/or abstract because they were a conference abstract, reported on assessment tools for medical technical skills, or because validation was performed within other medical specialties. The exclusion of these studies resulted in 115 studies of which the full article has been assessed by at least two independent reviewers. Of these studies, 31 were excluded since they did not report on validated assessment tools, 25 did not focus on teamwork performance, 28 concerned another medical specialty, and 26 were only a conference abstract. We have sought the opinion of a third reviewer for four of these articles [5,16–18], of which three were excluded [16–18]. Furthermore, we contacted six authors by e-mail, to verify whether the inclusion criteria were met. Three authors did not reply (including one reminder), one provided us with two published articles on the subject, one replied that there was no additional information, and one research group responded that their manuscript is submitted, but unfortunately not yet published. This resulted in five studies which were appropriate for inclusion, reporting on six assessment tools [5,19–22].

Included assessment tools

The six assessment tools that were developed and validated for the evaluation of obstetric teamwork performance in a simulated setting are: Clinical Teamwork Scale (CTS) [19], Human Factors Rating Scale (HFRS) [20], Global Rating Scale (GRS) [20], Assessment of Obstetric Team Performance (AOTP) [21,22], Global Assessment of Obstetric Team Performance (GAOTP) [21,22] and the Teamwork Measurement Tool (TMT) [5]. The characteristics of the assessment tools and the design of the validation process are presented in Table 2.

The most common teamwork behaviours included in the tools were: communication, situational awareness, leadership, and

Table 1
Accreditation Council for Graduate Medical Education (ACGME) Summary Recommendations.

Grading for the Overall Recommendation	
Class 1	The assessment method is ***recommended*** as a **core component** of the program's evaluation system.
Class 2	The assessment method ***can be considered for use*** as **one component** of the program's evaluation system.
Class 3	The assessment method ***can be used provisionally*** as **a component** of the program's evaluation process. Significant gaps in understanding of the assessment's value remain, so methods in this class are best suited for investigational research.
Criteria for Determining Level of Evidence	
Level A	Published data from methodologically sound evaluation studies of the method in multiple (more than 2) settings provides strong evidence for all components of the modified utility index (reliability, validity, ease of use, resources required, ease of interpretation, and educational impact).
Level B	Published data from methodologically sound evaluation studies of the method in a minimum of two settings provides some evidence of acceptable reliability and some evidence of validity and, ease of use, and educational impact. Acceptable evidence for ease of interpretation is available for methods used to make high-stakes decisions. Available evidence for ease of use and resources required suggests that the tool is usable by many programs.
Level C	Data from methodologically sound evaluation studies of the method provide evidence of acceptable reliability, validity, or educational impact. Available evidence for ease of use and resources required suggests that the tool is usable by many programs.

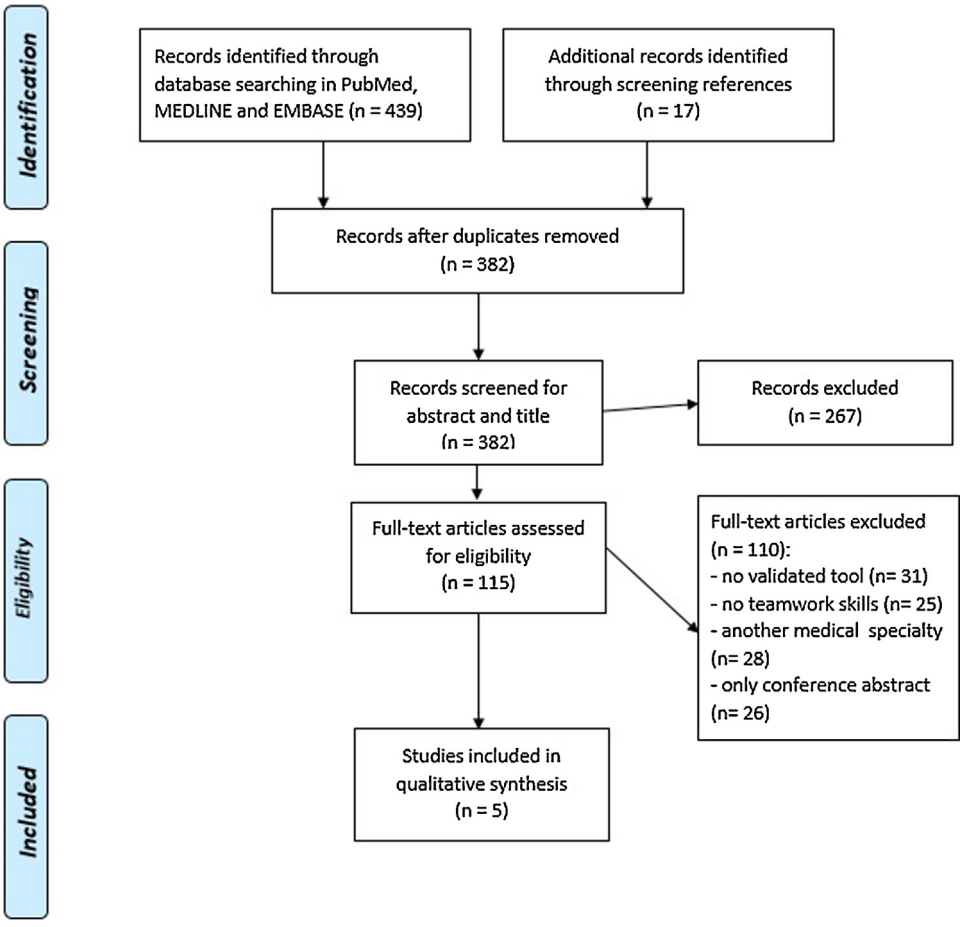

Fig. 1. Flow diagram demonstrating an overview of the selection process.

decision making. All studies included postgraduate healthcare professionals working within multi-professional obstetric care teams. Apart from the study by Siassakos et al. [5] and Guise et al. [19], the participating obstetric care teams were multi-disciplinary, meaning that their care teams also included anaesthesiologist and/or family medicine doctors. The AOTP and GAOTP were the

Table 2
Characteristics of included assessment tools.

Assessment Tool	Items	Type of item response	Medical specialities involved in teams	Setting for validation	Number of raters	Number of assessments used for validation
Clinical Teamwork Scale (CTS)	15	0-10 rating scale (and 1 Yes/No item)	Obstetrics	3 scripted simulated scenarios (with different predefined levels of performance)	3 raters	9 ratings
Teamwork Measurement Tool (TMT)	7	Multiple choice, counting of events, assigning of score	Obstetrics	19 simulated scenarios (1 clinical situation)	2 raters	38 ratings
Global Assessment of Obstetric Team Performance (GAOTP)	6	5 point Likert-scale	Obstetrics, anaesthesiology	12 simulated scenarios for usefulness (4 clinical situations) of which 3 were used for reliability measures	14 raters for usefulness; 3 raters for reliability measures	9 ratings for reliability measures; 168 ratings for usefulness
	6	5 point Likert-scale	Obstetrics, anaesthesiology, family medicine	136 simulated scenarios (4 clinical situations)	8 raters for reliability	1088 ratings for reliability measures
Assessment of Obstetric Team Performance (AOTP)	18	5 point Likert-scale	Obstetrics, anaesthesiology	12 simulated scenarios for usefulness (4 clinical situations) of which 3 were used for reliability measures	14 raters for usefulness; 3 raters for reliability measures;	9 ratings for reliability; 168 ratings for usefulness
	16	5 point Likert-scale	Obstetrics, anaesthesiology, family medicine	136 simulated scenarios (4 clinical situations)	8 raters for reliability measures	1088 ratings for reliability measures
Human Factors Rating Scale (HFRS)	45	5 point Likert-scale	Obstetrics, anaesthesiology	12 simulated scenarios (4 clinical situations)	9 raters for reliability measures	108 ratings for reliability measures
Global Rating Scale (GRS)	1	5 point Likert-scale	Obstetrics, anaesthesiology	12 simulated scenarios (4 clinical situations)	9 raters for reliability measures	108 ratings for reliability measures

Table 3
Psychometric and usefulness properties of the assessment tools.

Assessment tool	Validity	Reliability	Usefulness	ACGME Evidence grade & Overall recommendation	Oxford level of evidence
Clinical Teamwork Scale (CTS)	Construct validity: 60–82% of scores fell in the predefined teamwork level.	Inter-rater reliability, Pearson correlation coefficient: 0.94-0.96; ICC: 0.98. Overall agreement, Kappa: 0.78. Concordance, Kendall coefficient: 0.95. Largest variance due to rater – item interaction; variance due to the rater was low (0–0.30)	Completeness ranged from 89–100%. Accurateness: 12 items had 100% accuracies (±1 point), 3 items had accuracies of 67–89%. No tool-specific training, evaluators were familiar with crew resource management. Clear interpretation guidelines available.	C; Class 3	3b
Teamwork Measure-ment Tool (TMT)	Construct validity, Kendall's rank correlation, for: stating the emergency τ: −0.59; SBAR τ: 0.43; task allocation τ: 0.41; Room exits τ: −0.36; Situational awareness τ: 0.38; Supportive language τ: 0.44.	*Details not discussed*	*Details not discussed*	C; Class 3	3b
Global Assessment of Obstetric Team Performance (GAOTP)	*Details not discussed*	Internal consistency, Cronbach's α: 0.91. Inter-rater reliability, (single-rater) ICC: 0.34; (eight-rater) Cronbach's α: 0.81. Test-retest reliability, Pearson's correlation: 0.47.	All raters strongly agreed with ease of use. 13 out of 14 raters strongly agreed with ease of score interpretation.	C; Class 3	4
	Details not discussed	Pre-training internal consistency, Cronbach's α: 0.68. Post-training internal consistency, Cronbach's α: 0.87. Inter-rater reliability, pre-training ICC: 0.54; post-training ICC: 0.94.	Raters attended an 8 h workshop.		
Assessment of Obstetric Team Performance (AOTP)	*Details not discussed*	Internal consistency, pre-training Cronbach's α: 0.83; post-training Cronbach's α: 0.91	All raters strongly agreed with ease of use. Median amount of time spent was 7.5 min (1.5–50 min); time requirement evaluated as moderate and manageable by 75% of raters. 13 of 14 raters strongly agreed with ease of score interpretation. The effect of training was investigated.	C; Class 3	4
	Details not discussed	Internal consistency, Cronbach's α: 0.96. Test-retest reliability, Pearson's correlation: 0.47.	Raters attended an 8 h workshop.		
Human Factors Rating Scale (HFRS)	*Details not discussed*	Inter-rater reliability, (single-rater) ICC: 0.34; (nine-rater) Cronbach's α: 0.82	*Details not discussed*	C; Class 3	4
Global Rating Scale (GRS)	*Details not discussed*	Inter-rater reliability, (single-rater) ICC: 0.45; (nine-rater) Cronbach's α: 0.88	*Details not discussed*	C; Class 3	4

only tools that were studied in two settings, of which Tregunno et al. especially focused on the usefulness of these tools [21,22]. Five out of six tools used a Likert response scale with behavioural anchors [19–22].

Development of assessment tools

The used procedure for content selection was extensively described for the AOTP [21,22], GAOTP [21,22] and TMT[5]. For the development of the AOTP and GAOTP the researchers combined the following techniques to generate a list of themes and subthemes: narratives of teamwork behaviours by trainees, focus group sessions, literature reviews, statistical analysis methods, and expert opinions [21,22]. For the development of the TMT a literature study was performed to create a list of teamwork behaviours derived from evaluation studies [5], which was discussed and transformed by a multi-professional steering group. Subsequently, it was assessed whether the chosen teamwork behaviours were observable and measurable in video recordings of obstetric simulation-based scenarios. For the CTS, HFRS and GRS, the selection procedure of content was described briefly [19,20]. Regarding the item selection of the CTS, the authors only described that this was based on the principles of crew resource management [19]. The HFRS was adapted minimally from the Operating Room Management Attitudes' Questionnaire (ORMAQ) for the use in obstetrics. The specific adjustments and the rationale for these changes were not described. The same authors provided no information about the content development of the GRS [20].

Validity

Construct validity, meaning evidence that supports that the assessment tool is measuring the construct which is intended – i.e. teamwork performance – was determined for the CTS [19] and the TMT [5]. For the CTS, construct validity was assessed by comparing the distribution of scores, and raters' median scores, with the predefined teamwork level. Overall, they found that 60 to 82% of the scores fell within the predefined ranges. The median scores of all assessments (nine ratings) corresponded well with the predefined teamwork levels (100% within the predefined ranges). As a proof of construct validity, Siassakos et al. demonstrated significant correlations between several items on the TMT and specific categories of the Weller scale (a global rating scale which was validated using videotapes from an Anaesthesia Crisis Resource Management course) [23]. Content-related validity was described for the AOTP [22], GAOTP [22], and TMT [5]. Criterion-related validity was not reported in any of the studies.

Reliability

Reliability is considered in terms of internal consistency and reproducibility. Internal consistency measures whether items that intend to measure the same, have similar scores. It is based on the correlation of the scale's items [24]. One study, referring to two tools (AOTP and GAOTP), presented good internal consistency for both tools [21]. Reproducibility is defined as the stability of an instrument over time (test-retest) and inter-rater agreement at one point in time. Inter-rater reliability is determined for four assessment tools: good inter-rater reliability for the CTS [18] and poor inter-rater reliability for the HFRS, GRS and GAOTP [20,21]. Test-retest reliability was only determined for the AOTP and GAOTP, and appeared to be moderate [21]. All reliability measures are presented in Table 3.

Ease of use

All assessment tools were easily accessed and carried out. For one tool, the AOTP [22], the authors evaluated the time to complete the assessment tool. The median time to complete this tool was 7.5 min (ranged from 1.5 to 50 min), which was evaluated as moderate and manageable by the raters.

Resources required

No additional resources were required to use the assessment tools. However, the raters of the AOTP and GAOTP took part in an eight hour workshop [21]. For the CTS [19], GRS [20] and HFRS [20], this was not provided and the authors of the TMT [5] did not report on this item. Although all assessment tools could be used by individual raters, evidence for validity and/or reliability of some tools was based on the combination of scores obtained from more than one rater (TMT [5], GRS and HFRS [20]).

Ease of interpretation

All tools used an easy to understand scale. The CTS was the only assessment tool for which some form of normative data were presented [19]. The authors used a 10-point rating scale with three predefined levels of team performance which resulted in median scores for each level (poor/fair – average – good/perfect). Obtained scores can be compared with these predefined data.

Educational impact

Educational impact refers to the effect of the assessment tool on the performance of trainees or the curriculum program. As improvement of trainees' performance is a main reason to use an assessment tool, the ACGME guidelines recommend to evaluate its educational impact. However, none of the included studies investigated the educational impact.

Overall ACGME grading

By using the standards for tools and summary recommendations, an overall ACGME grading for each tool and a grading for each of the six topics were assigned (validity, reliability, ease of use, resources required, ease of interpretation, and educational impact; reported in Appendix B in Supplementary material) [12]. None of the studies received the highest ACGME grade (A) for validity nor reliability. All studies received the middle grade (B) for ease of use. The CTS is the only tool that provided evidence to justify the highest grade for resource required [19]. None of the tools reached the highest grade for ease of interpretation. Finally, none of the tools provided any information about educational impact.

All assessment tools received a level C of evidence (see Table 3). Although the CTS received the highest ACGME grades, the validation was limited to one study, which implies a level C of evidence. Consequently, a Class 3 overall recommendation was applied to all assessment tools which means they can be used provisionally as a component of a program's evaluation.

Risk of bias

All included studies were assessed according to the Oxford Centre for Evidence Based Medicine (OCEBM) levels of evidence for diagnostic studies (Table 3). Two studies [5,19] offered level 3b evidence, owing to non-consecutive cohorts. All other studies consisted of level 4 evidence due to poor or non-independent reference standards [20–22].

In all studies, teamwork performance was assessed by independent and blinded raters [5,19–22]. Morgan et al. (AOTP and GAOTP) reported blinding of the raters to the sequence of the training sessions. They additionally randomised the scenario order for each rater to minimise a learning effect [21]. For the validation process of the CTS, video recordings of teams were unidentified for the raters and viewed in no specific order [19]. The validation of the TMT included a randomised order and the authors were blinded to the site, timing, and type of training [5]. The authors of the HFRS and GRS did not report on blinding or randomisation of the videos used for validation [20].

Discussion

This systematic search and review provides an overview of the available validated assessment tools for the evaluation of obstetric teamwork performance in simulated settings. The included assessment tools resulted from a systematic literature search and were critically appraised for the quality of their psychometric and usefulness properties. This review demonstrates that several assessment tools for obstetric teamwork performance in simulated settings are available but the evidence supporting their psychometric properties remains quite limited.

We concluded that the Clinical Teamwork Scale by Guise et al. possesses the best psychometrics from the six included assessment tools in this review [19]. The authors were able to demonstrate a good reliability and validity, and the tool was easy to use. However, the educational impact of the assessment tool was not examined, similar to the validation of the other assessment

tools. Therefore, we assigned a C level of evidence and an overall recommendation Class 3 to the CTS. However, more research is desirable, including a larger number of raters and scenarios. More research is also needed for the TMT [5], especially to explore whether the TMT can distinguish between different levels of performance. Both studies received a level of evidence 3b, according to the Oxford Centre for Evidence Based Medicine (OCEBM).

With regard to the evaluation of assessment tools for teamwork performance in medical simulation settings, we encountered one other systematic review, which focused on hospital action teams [25]. They identified six tools for the assessment of team-centric communication in hospital action teams: Team Emergency Assessment Measure (TEAM) [26], Trauma Non-Technical Skills Scale (T-NOTECHS) [27], Simulation Team Assessment Tool (STAT) [28], Trauma Team Communication Assessment (TTCA), Communication and Teamwork Skills (CATS) [29], Observational Teamwork Assessment for Surgery (OTAS) [30]. A similar limitation of validity evidence, similar to our review, has been acknowledged by the authors [25]. The lack of information on the validity of assessment tools of teamwork performance is surprising, since training courses are increasingly focusing on the education of these skills. To ensure effective education of these skills, validated assessment tools are required. Of course one could argue that a subjective debriefing session is sufficient for providing a reliable and valid assessment. However, high stakes testing of trainees and obtaining robust evidence from educational research should profoundly rely on validated assessment methods. For the standardisation of the validation processes we recommend to use the ACGME guidelines [12].

The ongoing debate about standards for teamwork performance could be an explanation for the limited quality of the applied validation processes. The lack of a gold standard, which can be used to test construct validity, makes this even more complicated. Researchers use different sources to overcome this problem, for example the comparison with other validated teamwork assessment tools, and/or the use of expert opinions. Studies in which standards for obstetric teamwork performance are developed (such as Siassakos et al. [5]) provide a valid starting point for new assessment tools. Moreover, exploring other techniques for the assessment of obstetric teamwork performance might contribute to the development of a gold standard. This could, for example, include the moving behaviour of team members.

The strength of this review lies in the systematic way in which the quality of the available tools has been evaluated. The applied ACGME grading supplies the readers with a useful synopsis of available tools and guide them in choosing the most appropriate tool for their educational needs. Although we consider the application of the ACGME guidelines as an important strength of our study, a downside should be acknowledged. In practice, applying the grading rules was not always easy. This especially concerns the overall recommendation, as there were no explicit criteria described by the ACGME. Nonetheless, the two independent reviewers of the current study initially differed only about one study and easily reached consensus about the final recommendation. Despite this limitation of the grading rules, we found them very helpful in exploring the quality of the applied validation processes of the included assessment tools.

Several other limitations of our study should be acknowledged. In general, the quality of the applied validation processes was rather limited. This is important to keep in mind when interpreting the results of our review. Nevertheless, our review provides researchers with a useful guide for developing new, or validating existing, assessment tools. Secondly, we deliberately focused on the validation for simulated settings for obstetric teams. In the studies of Siassakos et al. [5] and Guise et al. [19], there were no multi-disciplinary teams included for the validation of assessment tools. Whenever these tools would be used in clinical practice, with multi-disciplinary teams, the psychometric characteristics could be different. However, with regard to the defined teamwork behaviours in the assessment tools, we do not expect any differences in clinical practice. Therefore, applying these tools in daily practice might be possible and could support constant improvement of teamwork performance. Finally, we focused only on obstetrics, since Bahl et al. have demonstrated the domain-specificity of teamwork behaviours in obstetrics [11], we can imagine that the same holds true for other specific domains (e.g. surgery, anesthesiology and pediatrics). We therefore recommend recognition of the domain-specific nature of teamwork behaviours.

In conclusion, several assessment tools for obstetric teamwork performance in simulated settings are available but the evidence supporting their psychometrics remains quite limited. Based on a systematic evaluation of the instruments' quality using the ACGME guidelines, we concluded that the Clinical Teamwork Scale [19] has the most comprehensive validation process, and the TMT [5] is a promising tool of which the reliability must further be investigated. Moreover, we recommend researchers to use the ACGME guidelines to support the design of future validation studies.

Conflicts of interest

All authors declare that they have no financial or personal relationships with other people or organisations that could inappropriately have influenced our work.

Funding

We received no funding for this study.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ejogrb.2017.06.034.

References

[1] Siassakos D, Fox R, Bristowe K, Angouri 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.
[2] Gordon M, Baker P, Catchpole K, Darbyshire D, Schocken D. Devising a consensus definition and framework for non-technical skills in healthcare to support educational design: a modified delphi study. Med Teach 2015;37:572–7.
[3] Martijn L, Jacobs A, Amelink-Verburg M, Wentzel R, Buitendijk S, Wensing M. Adverse outcomes in maternity care for women with a low risk profile in the netherlands: a case series analysis. BMC Pregnancy Childbirth 2013;13:219.
[4] Confidential Enquiry into Maternal Deaths in the United Kingdom (CMACE). Saving mothers' lives: reviewing maternal deaths to make motherhood safer: 2006–2008. the eight report. BJOG 2011;118:1–203.
[5] Siassakos D, Bristowe K, Draycott TJ, Angouri J, Hambly H, Winter C, et al. Clinical efficiency in a simulated emergency and relationship to team behaviours: a multisite cross-sectional study. BJOG 2011;118:596–607.
[6] Fransen AF, van de Ven J, Merién AE, de Wit-Zuurendonk L, Houterman S, Mol BW, et al. Effect of obstetric team training on team performance and medical technical skills: a randomised controlled trial. BJOG 2012;119:1387–93.
[7] Boet S, Bould MD, Fung L, Qosa H, Perrier L, Tavares W, et al. Transfer of learning and patient outcome in simulated crisis resource management: a systematic review. Can J Anaesth 2014;61:571–81.
[8] Cook DA, Hatala R, Brydges R, Zendejas B, Szostek JH, Wang AT, et al. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. JAMA 2011;306:978–88.
[9] Ericsson KA. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. Acad Med 2004;79:S70–81.
[10] McGaghie WC, Issenberg SB, Cohen ER, Barsuk JH, Wayne DB. Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence. Acad Med 2011;86:706–11.

[11] Bahl R, Murphy DJ, Strachan B. Non-technical skills for obstetricians conducting forceps and vacuum deliveries: qualitative analysis by interviews and video recordings. Eur J Obstet Gynecol Reprod Biol 2010;150:147–51.
[12] Swing SR, Clyman SG, Holmboe ES, Williams RG. Advancing resident assessment in graduate medical education. J Grad Med Educ 2009;1:278–86.
[13] van der Vleuten CP, Schuwirth LW. Assessing professional competence: from methods to programmes. Med Educ 2005;39:309–17.
[14] Fletcher GC, McGeorge P, Flin RH, Glavin RJ, Maran NJ. The role of non-technical skills in anaesthesia: a review of current literature. Br J Anaesth 2002;88:418–29.
[15] OCEBM Levels of Evidence Working Group*. The Oxford 2011 Levels of Evidence. Oxford Centre for Evidence-Based Medicine. http://www.cebm.net/index.aspx?o=5653 (Last accessed 14.06.2017).
[16] Balki M, Cooke ME, Dunington S, Salman A, Goldszmidt E. Unanticipated difficult airway in obstetric patients: development of a new algorithm for formative assessment in high-fidelity simulation. Anesthesiology 2012;117:883–97.
[17] Kim J, Neilipovitz D, Cardinal P, Chiu M. A comparison of global rating scale and checklist scores in the validation of an evaluation tool to assess performance in the resuscitation of critically ill patients during simulated emergencies (abbreviated as CRM simulator study IB). Simul Healthc 2009;4:6–16.
[18] Rovamo L, Mattila MM, Andersson S, Rosenberg P. Assessment of newborn resuscitation skills of physicians with a simulator manikin. Arch Dis Child Fetal Neonatal Ed 2011;96:F383–9.
[19] Guise JM, Deering SH, Kanki BG, Osterweil P, Li H, Mori M, et al. P. Osterweil, H. Li, M. Mori, et al., Validation of a tool to measure and promote clinical teamwork. Simul Healthc 2008;3:217–23.
[20] Morgan PJ, Pittini R, Regehr G, Marrs C, Haley MF. Evaluating teamwork in a simulated obstetric environment. Anesthesiology 2007;106:907–15.
[21] 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 simulationBMJ Qual Saf2120127882. BMJ Qual Saf 2012;21:78–82.
[22] Tregunno D, Pittini R, Haley M, Morgan PJ. Development and usability of a behavioural marking system for performance assessment of obstetrical teams. Qual Saf Health Care 2009;18:393–6.
[23] Weller J, Frengley R, Torrie J, Shulruf B, Jolly B, Hopley L, et al. Evaluation of an instrument to measure teamwork in multidisciplinary critical care teams. BMJ Qual Saf 2011;20:216–22.
[24] Downing SM. Reliability: on the reproducibility of assessment data. Med Educ 2004;38:1006–12.
[25] Rehim SA, DeMoor S, Olmsted R, Dent DL, Parker-Raley J. Tools for assessment of communication skills of hospital action teams: a systematic review. J Surg Educ 2016;74:341–51.
[26] Cooper S, Cant R, Porter J, Sellick K, Somers G, Kinsman L, et al. Rating medical emergency teamwork performance: development of the team emergency assessment measure (TEAM). Resuscitation 2010;81:446–52.
[27] Steinemann S, Berg B, DiTullio A, Skinner A, Terada K, Anzelon K, et al. Assessing teamwork in the trauma bay: introduction of a modified NOTECHS scale for trauma. Am J Surg 2012;203:69–75.
[28] Reid J, Stone K, Brown J, Caglar D, Kobayashi A, Lewis-Newby M, et al. The simulation team assessment tool (STAT): Development, reliability and validation. Resuscitation 2012;83:879–86.
[29] Frankel A, Gardner R, Maynard L, Kelly A. Using the communication and teamwork skills (CATS) assessment to measure health care team performance. Jt Comm J Qual Patient Saf 2007;33:549–58.
[30] Healey AN, Undre S, Vincent CA. Developing observational measures of performance in surgical teams. Qual Saf Health Care 2004;13:i33–40.