2019-09-25

Evaluation of teamwork assessment tools for interprofessional simulation: a systematic review

Wooding, E

10.1080/13561820.2019.1650730
Journal of Interprofessional Care
Taylor & Francis

Evaluation of teamwork assessment tools for interprofessional simulation: a systematic review

Accepted for publication in:

Journal of Interprofessional Care

https://doi.org/10.1080/13561820.2019.1650730

Wooding, E.L.[1,2], Gale, T.C.[1,3], Maynard, V.[1]

[1] Peninsula Medical School, Plymouth University

[2] Department of Paediatrics, Royal Devon and Exeter NHS Foundation Trust

[3] Department of Anaesthesia, University Hospitals Plymouth NHS Trust

Correspondence: Dr Tom Gale, Director of Assessment, Faculty of Medicine and Dentistry, Plymouth University, Drake Circus, Plymouth, United Kingdom PL4 8AA. E-mail: thomas.gale@plymouth.ac.uk Orcid ID https://orcid.org/0000-0003-4551-5860

Keywords: Teamwork, team effectiveness, interprofessional research, systematic review, education

Abstract

There is growing evidence supporting the use of simulation-based education to improve teamwork in the clinical environment, which results in improved patient outcomes. Interprofessional simulation improves awareness of professional roles and responsibilities, promotes teamwork and provides training in non-technical skills. Tools have been developed to assess the quality of teamwork during simulation, but the use of these tools should be supported by validity evidence in appropriate contexts. This study aims to assess the validity of teamwork tools used in simulation-based interprofessional training for healthcare workers and students, and to compare the design and reporting of these studies. Medline, EMBASE, ERIC and CINAHL were searched using terms synonymous with simulation, crew resource management, training, assessment, interprofessional, and team work, from 2007-2017. Interprofessional healthcare simulation studies involving objectively rated teamwork training were included. The initial search provided 356 records for review, of which 24 were ultimately included. Three tools demonstrated good validity evidence underpinning their use. However, three studies did not explore tool psychometrics at all, and the quality of reporting amongst these studies on design and participant demographics was variable. Further research to generate reporting guidelines and validate existing tools for new populations would be beneficial.

Introduction

Errors in healthcare cause significant patient morbidity and mortality and remain an important focus for clinical education, forming the basis for the development of Crew Resource Management (CRM) training in healthcare. Team Strategies and Tools to Enhance Performance and Patient Safety (TeamSTEPPS), a widely recognised CRM model (King, et al., 2008), was developed as a direct result of the Institute of Medicine report 'To Err is Human' (2000), which itself was inspired by industry including aviation. The importance of good teamwork in clinical care is well evidenced, and non-technical skills training, including in the form of simulation, plays a strong role in this (Gordon, Darbyshire & Baker, 2012). The literature abounds with examples of simulated learning interventions pertaining to improve teamwork or non-technical skills. In some cases, this is objectively measured by tools, but it is unclear to what extent these tools are fit for purpose and adequately validated (Rosen, et al., 2008; Onwochei, Halpern & Balki, 2017). This research will be relevant to those planning simulated teamwork training in the clinical environment or the skills laboratory for interprofessional groups.

Background

Rosen, et al., (2008) proposed a best practice framework for designing team performance measurement tools, proposing that teamwork training cannot be considered to be effective unless it is accurately measured and used to provide feedback to trainees on performance, areas for improvement and ongoing training needs. Key proposals pertinent to this review were that tools should be formed with an understanding of the theory underpinning teamwork

models and with a clear learning objective for the training. Tools should focus on observable behaviours and be used by trained observers from multiple sources, who are appraised by analysing tool performance (Rosen, et al., 2008).

On reviewing the literature around teamwork tools, it was noted that there were a wide variety of teamwork measurement tools available, but that the validity evidence supporting them was variable. A number of teamwork tools came up repeatedly in published literature, but in many cases were reused without consideration for the population studied and new tools were created despite validated tools existing for the relevant population. Previous systematic reviews have looked at the quality of teamwork assessment tools in Obstetrics (Onwochei, et al., 2017) and internal medicine (Havyer et al., 2014) and how simulation based interprofessional teamwork training changes behaviours or patient outcome (Fung, et al., 2015). Havyer, et al., (2016) carried out a systematic review focussing on teamwork assessment instruments in undergraduate education, some of which was interprofessional and relevant to this review. Following literature review it remained unclear which tools were best for assessing interprofessional teamwork simulation across all clinical specialities, amongst both under- and postgraduate healthcare professional groups.

Objectives

The objective of this review was to assess the validity of teamwork tools used in simulation-based interprofessional teamwork training for healthcare professionals and students. The review aimed to compare the design and reporting of studies in this area.

Methods

The protocol was designed by E.W. and revised by all authors (E.W., T.G., V.M.). A search was performed to locate studies in English on four databases: Medline, EMBASE, ERIC and CINAHL, using the following search terms and combination: simulat* AND ("cr* resource management" OR training OR education) AND assess* AND (inter*profession* OR multi*profession* OR multi*disciplin*) AND (team*work OR team train*). The databases were selected due to their education, interprofessional and medical foci. Each database was searched for studies published between 1 January 2007 and 31 March 2017. Database searching was supplemented with hand searching of key journals and sourcing additional papers from reference lists of included studies and relevant review articles.

Eligibility Criteria

To be eligible for inclusion studies needed to contain interprofessional groups of participants, either as qualified practitioners or students, taking part in simulated training with an intention to improve teamwork in healthcare. This teamwork had to be formally assessed in the study and an appraisal made of its change as a result of the intervention by raters who themselves were not participating in the simulation. Assessment of teamwork did not necessarily need to be the primary aim of the study, so long as it was appropriately assessed using a teamwork metric. Studies were excluded if they did not adequately describe their methodology for performing the intervention or analysing its effect on the teamwork. Interprofessional student groups were included, but interdisciplinary and intraprofessional learning alone was excluded.

Studies where the team entirely comprised non-clinical interprofessionals were excluded, but where these participants accompanied frontline clinical professionals, the studies were included. Simulated interventions from both the skills laboratory and clinical settings were eligible. The review is reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Statement, and the quality of studies was assessed using the PRISMA checklist and was a primary focus of this review (Moher, Liberati, Tetzlaff, & Altman, 2010).

Selection and Analysis

Titles and abstracts of identified citations were reviewed by E.W. Full text articles were reviewed by E.W. and areas of uncertainty regarding eligibility were discussed and resolved by consensus by E.W., T.G. and V.M. One researcher (E.W.) extracted data and entered it onto a spreadsheet for evidence synthesis. Data extraction incorporated study design, location, participant profile, course design, outcome measures, teamwork measurement tool and its psychometrics, rater credentials and follow-up outcome measurement. Literature search identified 233 articles, and a further 123 were added through hand searches and reference lists. 19 duplicates were removed, and 328 citations were screened. 89 full text articles met screening criteria and were reviewed in full, providing 24 articles which met all inclusion criteria and contributed to this review. The 65 remaining full text articles were removed due to lack of eligible outcomes (n = 21), uniprofessional groups of participants (n = 11), no objective rater of teamwork (n = 15), or ineligible educational intervention (n = 18). This process is summarised in Figure One. The data were analysed and reported using a narrative description as meta-analysis

was deemed unsuitable for the dataset due to the heterogeneity of study designs.

Included studies were evaluated according to Downing's validity framework for evidence of validity of each tool in specific contexts. (Downing, 2003) Validity evidence may commonly include test re-test reliability, reproducibility and generalisability of scores, and statistical characteristics of assessment components. Reliability of the teamwork tool expresses the extent to which the results can be reproduced. This can include test re-test reliability and interrater agreement. Interrater agreement is high when multiple raters assign a consistent score to each other when using the same assessment tool (Downing, 2003). Interrater agreement may be expressed using the intraclass correlation coefficient (ICC), whilst Cronbach's alpha measures internal consistency (Onwochei et al., 2017). Test re-test reliability looks at the extent to which an assessment tool would produce the same test score if applied on multiple occasions (Berchtold, 2016).

Results

Of the 356 records identified on database and hand searches, 19 studies from 24 papers met all inclusion criteria and contributed to this review (Auerbach, et al., 2014; Bradley, Cooper, & Duncan, 2009; Burton, et al., 2011; Cooper, et al., 2010; Daniels, Lipman, Harney, Arafeh, Druzin., 2008; Frengley, et al., 2011; Ghazali, et al., 2016; Hobgood, et al., 2010; Jankouskas, et al., 2007; MacDonnell, Rege, Misto, Dollase, George., 2012; Morgan, Pittini, Regehr, Marrs, Haley, 2007; Morgan, et al., 2012; Oriot, Bridier, & Ghazali, 2016; Paige, et al., 2014; Patterson, Geis, Falcone, LeMaster, & Wears, 2013; Phitayakorn, Minehart, Hemingway, Pian-Smith,

Petrusa, 2015; Rovamo, Nurmi, Mattila, Suominen, Silvennoinen, 2015; Sigalet, et al., 2013; Sigalet, Donnon, & Grant, 2015; Walker, et al., 2011; Weller, et al., 2011; Weller, et al., 2013; Zhang, Miller, Volkman, Meza & Jones, 2015). In the case of three studies, two or more manuscripts were published for each utilising the same dataset. For the purpose of this review these were considered to be three studies, rather than seven. The first was Oriot, et al. (2016) and Ghazali, et al. (2016), the second set was Weller, et al. (2011), Weller, et al. (2013), and Frengley, et al. (2011). The final pair was Sigalet, et al. (2013), and Sigalet, et al., (2015). This made the total number of studies included as 19.

Study Characteristics and Demographics

The studies included spanned from 2007 to 2016. Five studies employed comparator groups (Bradley, et al., 2009; Hobgood et al., 2010; Morgan, et al., 2012; Oriot, et al., 2016; Rovamo, et al., 2015). Oriot, et al. (2016) and Ghazali, et al. (2016) in their study used two comparator groups, but the control group was not interprofessional. 14 studies did not use comparator groups or controls (Auerbach, et al., 2014; Burton, et al., 2011; Cooper, et al., 2010; Daniels, et al., 2008; Jankouskas, et al., 2007; MacDonnell, et al., 2007; Morgan et al., 2007; Paige, et al., 2014; Patterson, et al., 2013; Phitayakorn, et al., 2015; Sigalet, et al., 2013, 2015; Walker, et al., 2011; Weller, et al., 2011, 2013; Zhang, et al., 2015).

13 studies stated the exact total number of participants attending their simulations (Auerbach, et al., 2014; Bradley, et al., 2009**;** Burton, et al., 2011; Daniels, et al., 2008; Hobgood, et al., 2010; Jankouskas, et al., 2007**;** Oriot, et al., 2016; Morgan, et al., 2007; Patterson, et al., 2013;

Phitayakorn, et al., 2015; Rovamo, et al., 2015; Sigalet, et al., 2013, 2015; Zhang, et al., 2015). This totalled 1740 people. The remaining, Morgan, et al. (2012) define number of 'performances' rather than number of individual participants, stating 136 performances completed by 10 teams. Three studies stated number of simulations or teams, but not specific participant numbers (Cooper, et al., 2010; Walker, et al., 2011; Weller, et al., 2011).

Eight studies cited demographic data on participants. Bradley, et al., (2009) collected detailed information on participants covering baseline characteristics, prior qualifications and prior experience, including interprofessional learning, teamwork or leadership experience and resuscitation training. Morgan, et al. (2007) gathered baseline data and previous experience in simulation, but also asked participants to rate their level of sleep deprivation and stress levels at the time of participation. Rovamo, et al. (2015) collected data on the clinical and academic experience of participants in relation to the clinical content of the simulation, and years of working experience, but not on non-technical skills knowledge or training, or further demographics such as age or gender. Sigalet, et al. (2013) report 82% female participants and collected data on previous team-based learning, experienced by a 93% proportion of participants. They specifically asked about prior interprofessional learning, reported by 11.7% of the total 196 participants. Daniels, et al. (2008) reported number of years of postgraduate experience demonstrated by participants and years of experience in labour and delivery practice, but no further demographics were given. Paige, et al. (2014) cited ethnic origin and clinical role of participants, and that gender distribution was even, but without further detail. The remaining 11 studies lacked demographic data on participants.

Team Profiles

Only studies describing interprofessional simulation were included. Of these, 12 study teams contained nurses, and 11 contained doctors. Other professions stated included midwives (Rovamo, et al., 2015), respiratory therapists (Burton, et al., 2011; Patterson, et al., 2013), scrub technicians (Phitayakorn, et al., 2014), paramedics (Oriot, et al., 2016; Patterson, et al., 2013) and physician assistants (Auerbach, et al., 2014). Other studies included interprofessional student groups, either exclusively, in the case of Zhang, et al., (2015) with exclusively physical therapy and nursing students; respiratory therapy, nursing and medical students by Sigalet, et al., (2013), medical, nursing and pharmacy students in MacDonnell, et al., (2012), and five studies with medical and nursing students only (Cooper, et al., 2010; Bradley, et al., 2009; Hobgood, et al., 2010; Paige, et al., 2014). Auerbach, et al., (2014) was the only study to state inclusion of both professionals and students, in this case medical students. They also detailed a number of other staff groups involved in simulations, namely allied health workers, social workers, diagnosticians and transport workers. However, the extent of their role in the simulations was not fully elucidated. This is also true of Patterson, et al., (2013) who stated the presence of patient care assistants and "others" in their simulations. Seven studies specifically named the presence of interdisciplinary professionals within their interprofessional group. (Daniels, et al., 2008; Jankouskas, et al., 2007; Morgan, et al., 2007; Morgan, et al., 2012; Phitayakorn, et al., 2015; Rovamo, et al., 2015; Walker, et al., 2011).

Simulation Context

All but one study (Hobgood, et al., 2010) stated the clinical context of the simulation scenarios. The most common content was paediatric emergencies (Auerbach, et al. 2014; Jankouskas, et al., 2007; Oriot, et al., 2016; Patterson, et al., 2013; Sigalet, et al., 2013) and neonatal emergencies (Burton, et al., 2011; Rovamo, et al., 2015; Sigalet, et al., 2013). Four studies focused on anaesthetic or surgical complications mid-operation (Daniels, et al., 2008; Paige, et al., 2014; Phitayakorn, et al., 2015; Morgan, et al., 2012), three on trauma (Auerbach, et al., 2014; Paige, et al., 2014; Zhang, et al., 2015), and three on obstetric emergencies (Daniels, et al., 2008; Morgan, et al., 2007; Morgan, et al., 2012). Four studies focused primarily on resuscitation skills (Bradley, et al., 2009; Cooper, et al., 2010; Walker, et al., 2011; Weller, et al., 2011), although most studies had a resuscitation element.

Simulations in 14 of the 19 studies took place in the simulation suite (Bradley, et al., 2009; Burton, et al., 2011; Daniels, et al., 2008; Ghazali, et al., 2016; Hobgood, et al., 2010; Jankouskas, et al., 2007; MacDonnell, et al., 2012; Morgan, et al., 2007; Morgan, et al., 2012; Paige, et al., 2014; Patterson, et al., 2013; Rovamo, et al., 2015; Sigalet, et al., 2013; Zhang, et al., 2015), two took place solely *in situ*, i.e. in the clinical environment (Auerbach, et al., 2014; Phitayakorn, et al., 2015), and in Weller, et al. (2011) it was unclear whether the 'recreated ICU' was in the simulation suite or *in situ*. Two studies incorporated both simulation suite and *in situ* elements (Cooper, et al., 2010; Walker, et al., 2011).

Tool Selection

10 studies used a previously developed tool in its original state (Burton, et al., 2011; Cooper, et al., 2010; Jankouskas, et al., 2007; Morgan, et al., 2012; Paige, et al., 2014; Patterson, et al., 2013; Phitayakorn, et al., 2015; Rovamo, et al., 2015; Sigalet, et al., 2013; Zhang, et al., 2015). In the case of Morgan et al., (2012) the study was created by the same group and previously validated (Tregunno, et al., 2008) These were the Objective Teamwork Assessment System (OTAS) (Phitayakorn, et al., 2015), the Team Emergency Assessment Measure (TEAM) (Cooper, et al., 2010; Rovamo, et al., 2015), the KidSIM Team Performance Scale (Sigalet, et al., 2013) and the Anaesthetists' Non-Technical Skills (ANTS) behavioural scale (Jankouskas, et al., 2007; Patterson, et al., 2013). Hobgood, et al., (2010), and Weller, et al., (2013) also used the Mayo Scale, but adapted it for their use.

Four groups developed their own tool for the purpose of their study. (Auerbach, et al., 2014; Daniels, et al., 2008; MacDonnell, et al., 2012; Oriot, et al., 2016). The Team Average Performance Assessment Scale (TAPAS) developed by Oriot, et al., (2016) is a 129-item scale in six sections designed for assessing teamwork in paediatric and life-threatening emergencies. Their tool design was linked to anticipated learning outcomes and developed by subject experts, although it is unclear whether development was also informed by relevant literature. Daniels, et al., (2008) developed a Checklist of Expected Actions, incorporating both anticipated clinical outcomes and a behavioural performance domain rated on a Likert scale, which was described as informed by contemporary literature. Two studies did not name their teamwork measures but described them as a validated tool (Auerbach, et al., 2014), or a tool based on a

validated tool (MacDonnell, et al., 2007) without detailing specific psychometrics or how the tool was informed or developed.

Quality of teamwork rating process

Included studies had objective ratings performed during or after the simulation to assess the quality of teamwork observed. Six studies used interprofessional groups of raters commensurate with the populations studied (Cooper, et al., 2010; Jankouskas, et al., 2007; Morgan, et al., 2012; Phitayakorn, et al., 2015; Sigalet, et al., 2013; Zhang, et al., 2015). A further four studies used doctors alone as raters (Daniels, et al., 2008; Oriot, et al., 2016; Rovamo, et al., 2015; Weller, et al., 2011). Others did not state the profession of their rater, although one was the lead researcher in the study (Auerbach, et al., 2014), two described "trained" raters (Paige, et al., 2014; Patterson, et al., 2013), and as "independent scorers" (Hobgood, et al., 2010). Only one study (Burton, et al., 2011) did not provide any information on the raters.

Training provided for raters was variable and inconsistently reported. However, nine studies described some form of rater training in the teamwork tool utilised (Burton, et al., 2011; Daniels, et al., 2008; Hobgood, et al., 2010; Morgan, et al., 2012; Oriot, et al., 2016; Paige, et al., 2014; Patterson, et al., 2013; Phitayakorn, et al., 2015; Weller, et al., 2011). No specific rater training was cited in four studies (Auerbach, et al., 2014; Rovamo, et al., 2015; Sigalet, et al., 2013; Zhang, et al., 2015), although Zhang, et al. (2015) did choose raters with previous training in teamwork skills.

Video rating was predominantly performed retrospectively, in 14 cases (Bradley, et al., 2009; Daniels, et al., 2008; Burton, et al., 2011; Hobgood, et al., 2010; Jankouskas, et al., 2007; Morgan, et al., 2007; Morgan, et al., 2012; Patterson, et al., 2013; Phitayakorn, et al., 2015; Rovamo, et al., 2015; Sigalet, et al., 2013; Walker, et al., 2011; Weller, et al., 2011; Zhang, et al., 2015). Contemporaneous rating took place in four cases (Auerbach, et al., 2014; Oriot, et al., 2016; MacDonnell, et al., 2011; Paige, et al., 2014). One study used a mixture of contemporaneous and retrospective rating (Cooper, et al., 2010).

Tool Psychometrics

Four studies used tools where their psychometric properties had been previously investigated and demonstrated to be acceptable (Burton, et al., 2011; Jankouskas, et al., 2007; Patterson, et al., 2013; Rovamo, et al., 2015), although it was not specifically clarified that the validity has been demonstrated for the population in question. Seven studies (Cooper, et al., 2010; Morgan, et al., 2007; Morgan, et al., 2012; Oriot, et al., 2016; Walker, et al., 2011; Weller, et al., 2011; Zhang et al., 2015) assessed the reliability and validity of their tools during their studies.

Hobgood, et al. (2010) modified their tool and found acceptable intraclass correlations on 19 out of 20 items on their tool, and also analysed variance but did not extensively re-calculate the psychometrics. Several studies did not refer to, or attempt to calculate, the psychometrics of the tool employed (Auerbach, et al., 2014; Daniels, et al., 2008; MacDonnell, et al., 2012)

making it difficult to draw satisfactory conclusions from their findings.

Cronbach's alpha was calculated in six studies (Cooper, et al., 2010; Morgan, et al., 2007; Morgan, et al., 2012; Oriot, et al., 2016; Rovamo, et al., 2015; Sigalet, et al., 2013; Walker, et al., 2011; Weller, et al., 2011). Intra-class correlation and/or correlation co-efficients were calculated in six studies (Burton, et al., 2011; Hobgood, et al., 2010; Morgan, et al., 2007; Oriot, et al., 2016; Paige, et al., 2014; Walker, et al., 2011), and interrater reliability in three (Morgan, et al., 2012; Phitayakorn, et al., 2015; Rovamo, et al., 2015). However, poor to moderate interrater reliability was demonstrated by Rovamo, et al. (2015). See Table One for further details of statistical significance.

Outcome Measures

Six studies aimed to validate a teamwork tool as their primary objective. (Cooper, et al., 2010; Morgan, et al., 2007; Morgan, et al., 2012; Oriot, et al., 2016; Walker, et al., 2014; Weller, et al., 2011). Zhang, et al. (2015) specifically aimed to improve the subjectivity of their tool, the TPOT, using targeted behavioural markers. 13 studies assessed quality of teamwork as a surrogate measure of performance or as a secondary aim (Auerbach, et al., 2014; Bradley, et al., 2009; Burton, et al., 2011; Daniels, et al., 2008; Hobgood, et al., 2010; Jankouskas, et al., 2007; Morgan, et al., 2007; Oriot, et al., 2016; Paige, et al., 2014; Patterson, et al., 2013; Phitayakorn, 2015; Rovamo, et al., 2015; Sigalet, et al. 2013). Three studies specifically measured the impact of simulation on teamwork scores over other media (Hobgood, et al., 2010, Rovamo, et al., 2015, Sigalet, et al., 2013).

Follow up

No studies carried out specific follow up cohorts, but some studies continued simulation over a long period of time and noted an improvement in teamwork scores over time (Auerbach, et al., 2014; Burton, et al., 2011; Sigalet, et al., 2013; Weller, et al., 2011).

Discussion

The importance of teamwork in the clinical setting is undeniable, however there is a lack of agreement as to which tools are best for assessing teamwork in interprofessional simulation settings. All assessment methods in clinical education should be underpinned by validity evidence, without which they lack inherent meaning. This is no less relevant to the assessment of teamwork in interprofessional simulation. The following tools were used or adapted for use in the included studies: the Objective Teamwork Assessment System (OTAS) (Phitayakorn, et al., 2015), the Team Emergency Assessment Measure (TEAM) (Cooper, et al., 2010; Rovamo, et al., 2015), the KidSIM Team Performance Scale (Sigalet, et al., 2013) and the Anaesthetists' Non-Technical Skills (ANTS) behavioural scale (Jankouskas, et al., 2007; Patterson, et al., 2013), the Mayo Scale (Hobgood, et al., 2010; Weller, et al., 2011), the Observational Skill-based Clinical Assessment tool for Resuscitation (OSCAR) (Walker, et al., 2011), The Team Average Performance Assessment Scale (TAPAS) (Oriot, et al., 2016), the Team Performance Observation Tool (TPOT) (Zhang, et al., 2015), the Assessment of Obstetrical Team Performance (AOTP) and Global AOTP, (Morgan, et al., 2012) and a Checklist of Expected Actions. (Daniels, et al., 2008) One included study (Auerbach, et al., 2014) used a validated, but unnamed tool. With

such a variety of teamwork tools being utilised, it is important to consider which of the teamwork domains are actually being measured. Teamwork can be sub-divided into communication, situation monitoring, leadership and mutual support based on the Team Strategies and Tools to Enhance Performance and Patient Safety (TeamSTEPPS) principles, which are widely recognised (King, et al., 2008, p.10). However, the content of tools and definitions of teamwork are highly subjective and variable. The problem of loose definitions between studies in non-technical skills training is corroborated in a related systematic review (Gordon, et al., 2012).

Some tools, such as the ANTS had previously tested psychometrics, but they were not validated for the specific new population. (Jankouskas, et al., 2007; Patterson, et al. 2013) Where psychometric properties of tools were explored, these generally demonstrated acceptable to good internal consistency (e.g. Burton, et al., 2011; Cooper, et al., 2010; Hobgood, et al., 2010; Morgan, et al., 2007; Morgan, et al., 2012; Oriot, et al., 2016; Paige, et al., 2014; Phitayakorn, 2015; Rovamo, et al., 2015; Sigalet et al., 2013; Walker, et al., 2011; Weller, et al., 2011; Zhang, et al., 2015) Some studies did not explore tool psychometrics (Auerbach, et al., 2014; Daniels, et al., 2008; MacDonnell, et al., 2012) Best practice would be to maximise the validity evidence supporting the use of all teamwork tools ensuring where pre-existing validation had taken place that it was specifically applicable to the current study set up. Where tools were adapted from their original format, it was not entirely clear whether these adaptations were necessary or completed appropriately. Of those studies that developed their own teamwork tools, none fully met the best practice guidance for team performance measurement developed by Rosen, et al.

(2008). Daniels, et al., (2008) informed the development of their tool by referring to current literature. However, it was not clear whether the tool development was also linked to the learning objectives of the team training. Oriot, et al., (2016) linked their tool design to learning outcomes for training and development was guided by experts.

In this review, those tools with the strongest validity evidence supporting them were the TEAM tool (Cooper, et al., 2010; Rovamo, et al., 2015), TPOT (Zhang, et al., 2015) and GATOP/AOTP (Morgan, et al., 2012). They all demonstrated very good internal consistency with Cronbach's alpha values >0.89 obtained from samples sized greater than 72. Zhang, et al.'s TPOT (2015) demonstrated that higher TPOT scores were associated with fewer team errors ($p=0.0008$), with good test re-test reliability ($k=0.707$, $p<0.001$) and good interrater reliability ($k=0.73$). The TPOT was a tool created as part of the TeamSTEPPS curriculum. It has been validated elsewhere for nursing teams, similarly demonstrating good internal consistency (Maguire, Bremner, & Yanosky, 2014). The paper describing the initial development of the AOTP and GAOTP was published by Tregunno, Pittini, Haley, and Morgan (2009). The same study group published on its validation in 2012 (Morgan, et al.). Whilst Morgan, et al. (2012) demonstrated good tool psychometrics our systematic review did not demonstrate the tool's wider use in Obstetrics at present. The TEAM tool performed moderately well in terms of the rater index of agreement (0.41) and poor to moderate inter-rater reliability. (Rovamo, et al., 2015) A better interrater reliability of 0.55 was demonstrated by Cooper, et al. (2010) for the same tool, which is considered fair. The TEAM tool has been extensively validated in other interprofessional studies, including in live clinical resuscitations and *in situ* simulations (Cooper, et al., 2016;

Maignan, et al., 2016). Maignan, et al. (2016) demonstrated very good psychometrics, however their simulations were rated by simulation participants, and as such was excluded from this review. The TEAM tool demonstrates promise as a reliable and valid teamwork tool and merits further validation for different settings. The Mayo scale was used by three earlier studies included in this review, when it was newest and most popular (Burton, et al., 2011; Hobgood, et al., 2010; Weller, et al., 2011). However, the Mayo scale was validated initially for use as a self-rating rather than objective scale (Malec, et al., 2007). Limited psychometrics were explored by Hobgood, et al. (2010) with an acceptable intraclass correlation coefficient, and Burton, et al., (2011) with moderate reviewer correlation (Pearson's coefficient=0.41, $p<0.001$).

The validity evidence supporting the teamwork tools included in this review predominantly assess internal consistency using Cronbach's alpha, either alone (Sigalet, et al., 2013) or in combination with one other validity measure, such as inter-rater reliability (Morgan, et al., 2012; Rovamo, et al., 2015; Walker, et al., 2011) or intraclass correlation coefficient (Oriot, et al., 2016). The most detailed explorations of validity evidence were performed on the studies assessing the TEAM scale, especially Cooper, et al., (2010) who assessed construct, content and concurrent validity, as well as internal consistency. The Mayo Scale studies collectively assess internal consistency, construct validity and intraclass correlation coefficient. (Burton, et al., 2011; Hobgood, et al., 2010; Weller, et al., 2011). Zhang, et al., (2015) assessed the test re-test reliability, interrater reliability and internal consistency of the TPOT. Future studies evaluating team work tools should go beyond measuring interrater reliability and should make an attempt to include other measures of validity evidence such as construct validity and internal

consistency to investigate whether the tool is providing an accurate measure of appropriate constructs within each setting.

The quality of teamwork scoring can be improved by using multiple, trained raters. In the included studies the quality of rating, and rater training, was generally appropriate or very good. The majority of rating took place retrospectively with trained raters. Only six studies explicitly used interprofessional rater teams, which would be most appropriate for the population studied raters (Cooper, et al., 2010; Jankouskas, et al., 2007; Morgan, et al., 2012; Phitayakorn, et al., 2015; Sigalet, et al., 2013; Zhang, et al., 2015). The most comprehensive rater training and rating process and training was provided by Morgan, et al. (2012), where eight interprofessional reviewers attended an eight hour workshop on rating, before all blindly reviewing 136 simulation performances. In some studies rating was carried out by individuals who were simulation faculty or the primary researcher, (Auerbach, et al., 2014; Bradley, et al., 2009; Daniels, et al., 2008; MacDonnell, et al., 2012), or the status of the rater was not stated (Burton, et al., 2011; Hobgood, et al., 2010; Paige, et al., 2014; Patterson, et al., 2013; Walker, et al., 2011). Independent rating by adequately trained raters would be preferable, an assertion reflected in another recent systematic review (Onwochei, et al., 2017). More detailed descriptions of rater demographics, experience and training would be beneficial to draw more detailed conclusions on quality of training and its relationship with the rating and tool validation process. Havyer, et al. (2013) notes that whilst many studies purport to measure teamwork the situations within the study are somewhat contrived team situations. To adequately measure teamwork, even in a simulated environment, more realistic

representations of interprofessional collaboration, and linking teamwork measurement to patient outcome to a greater extent, would make study findings more clinically relevant and meaningful. In undergraduate teams it would also be beneficial to begin measuring teamwork skills pre-clinically, to explore how this changes through time and to reinforce the benefits of interprofessional collaboration at the earliest juncture. (Havyer, et al., 2016) Whittaker, Abboudi, Khan, Dasgupta, Ahmed, (2015) in their systematic review of teamwork assessment tools in surgery was not eligible for inclusion in this review as its focus was not on simulation in interprofessional groups. However, individual studies included a focus on these themes, albeit not simultaneously. Whilst it does not meet the criteria for our review, it provides a useful overview of teamwork tools used in a surgical context. Our study adds to the literature as it focusses on the measurement of interprofessional teamwork in simulation across both professional and student populations, and all specialties.

Limitations of review

The limitations of this review can be sub-divided into those related to the studies included and those related to the review process. Publication bias is such that studies with lower levels of impact did not reach publication and were therefore excluded. Incomplete descriptions of methodology by included studies limited our ability to interpret and synthesise findings. The study period was for 10 years from 2007-2017. However, there is a predominance of North American studies (n=11). Additionally, only eight studies cited full demographic data. Location and lack of participant demographics affects the generalisability of findings.

In terms of our review process, we included all studies which claimed to measure teamwork, providing they met all other inclusion criteria, without enforcing a strict definition of what that teamwork involved. Simulations which took place in educational and clinical settings were included and deemed comparable. We also included teams including any groups of clinical interprofessionals from all specialties, which may affect applicability of findings.

Conclusion

This review aimed to report on the validity of teamwork tools used to objectively assess interprofessional simulation teamwork training for healthcare professionals and students, and on the design and quality of reporting of these studies. We summarised 19 studies. The strongest psychometrics were reported by the TEAM tool (Cooper, et al., 2010; Rovamo, et al., 2015), TPOT (Zhang, et al., 2015) and the GATOP/AOTP (Morgan, et al., 2012). The methodological quality of studies was mostly reasonable, however reporting of the details of specific interventions was poor. Reporting of tool design in line with best practice reporting guidance, (Rosen, et al., 2008) where relevant, was limited. Where new tools have been generated, their psychometric properties were often not adequately explored or inferior to the validation of existing tools such as TPOT or TEAM. Further validation of these tools in new interprofessional settings, and in similar settings with improved methodologies, would be beneficial to underpin their use. Where existing tools are re-used to assess teamwork, they should be chosen on the basis of their validity for the population studied or re-validated for that population. When assessing the psychometric properties of tools researchers should extend beyond assessment of internal consistency alone to consider the construct validity and

internal consistency. Generating a framework for the reporting of studies assessing teamwork in simulation could improve the methodological quality of future studies and is a possible area for future research.

References

Auerbach, M., Roney, L., Aysseh, A., Gawel, M., Koziel, J., Barre, K., … Santucci, K. (2014) In situ pediatric trauma simulation: assessing the impact and feasibility of an interdisciplinary pediatric in situ trauma care quality improvement simulation program. Pediatric Emergency Care, 30, 884-891.

Berchtold, A. (2016) Test-retest: agreement or reliability? (2016) Methodological Innovations, 9, 1-7. (doi:10.1177/2059799116672875).

Bradley, P., Cooper, S., & Duncan, F. (2009) A mixed-methods study of interprofessional learning of resuscitation skills. Medical Education, 43, 912.922. (doi:10.1111/j.1365-2923.2009.03432.x).

Burton, K.S., Pendergrass, T.L., Byczkowski, T.L., Taylor, R.G., Moyer, M.R., Falcone, R.A., … Geis, G.L. (2011) Impact of simulation-based Extracorporeal Membrane Oxygenation training in the simulation laboratory and clinical environment. Simulation in Healthcare, 6, 284-291. (doi:10.1097/SIH.0b013e3182dfcea).

Cooper, S., Cant, R., Porter, J., Sellick, K., Somers, G., Kinsman, L., … Nestel, D. (2010) Rating medical emergency teamwork performance: development of the team emergency assessment measure (TEAM). Resuscitation, 81, 446-452. (doi:10.1016/j.resuscitation.2009.11.027).

Daniels, K., Lipman, S., Harney, K., Arafeh, J., & Druzin, M. (2008) Use of simulation based team training for obstetric crises in resident education. Simulation in Healthcare, 3, 154-160. (doi:10.1097/SIH.0b013e31818187d9).

Downing, S.M. (2003) Validity: on the meaningful interpretation of assessment data. Medical Education, 37, 830-837.

Fletcher, G., Flin, R., McGeorge, P., Glavin, R., Maran, N., & Patey, R. (2003). Anaesthetists' non-technical skills (ANTS): evaluation of a behavioural marker system. British Journal of Anaesthesia, 90, 580-588.

Frengley, R.W., Weller, J.M., Torrie, J., Dzendrowskyj, P., Yee, B., Paul, A.M., … Henderson, K.M. (2011) The effect of a simulation-based training intervention on the performance of established critical care unit teams. Critical Care Medicine, 39, 2605–2611. (doi:10.1097/CCM.0b013e3182282a98).

Fung, L., Boet, S., Bould, M.D., Qosa, H., Perrier, L., Tricco, A., … Reeves, S. (2015) Impact of crisis resource management simulations-based training for interprofessional and interdisciplinary teams: a systematic review. Journal of Interprofessional Care, 29, 433-444. (doi:10.3109/13561820.2015.1017555).

Ghazali, D.A., Ragot, S., Breque, C., Guechi, Y., Boureau-Voultoury, A., Petitpas, F., … Oriot, D. (2016) Randomized controlled trial of multidisciplinary team stress and performance in

immersive simulation for management of infant in shock: study protocol. Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine, 24, 36. (doi:10.1186/s13049-016-0229-0).

Gordon, M., Darbyshire, D., Baker, P. (2012) Non-technical skills training to enhance patient safety: a systematic review. Medical Education, 46, 1042-1054. (doi:10.1111/j.1365-2923.2012.04343.x).

Havyer, R., Wingo, M., Comfere, N., Nelson, D., Halvorsen, A., McDonald, F., Reed, D. (2014) Teamwork Assessment in Internal Medicine: A Systematic Review of Validity Evidence and Outcomes. J GEN INTERN MED 29: 894-910. (https://doi.org/10.1007/s11606-013-2686-8)

Havyer, R., Nelson, D., Wingo, M., Comfere, N., Halvorsen, A., McDonald, F., Reed, D. (2016) Addressing the Interprofessional Collaboration Competencies of the Association of American Medical Colleges: A Systematic Review of Assessment Instruments in Undergraduate Medical Education. Academic Medicine 91, 865-888. (https://doi.org/10.1097/ACM.0000000000001053).

Hobgood, C., Sherwood, G., Frush, K., Hollar, D., Maynard, L., Foster, B., … Taekman, J. (2010) Teamwork training with nursing and medical students: does the method matter? Results of an interinstitutional, interdisciplinary collaboration. Quality and Safety in Health Care, 19, e25. (doi:10.1136/qshc.2008.031732).

Institute of Medicine. (2000) *To err is human: building a safer health system*. Washington, DC: National Academy Press. Retrieved from: http://www.nationalacademies.org/hmd/~/media/Files/Report%20Files/1999/To-Err-is-Human/To%20Err%20is%20Human%201999%20%20report%20brief.pdf

Jankouskas, T., Bush, M., Murray, B., Rudy, S., Henry, J., Dyer, A., Liu, W., Sinz, E. (2007) Crisis Resource Management: Evaluating Outcomes of a Multidisciplinary Team. Simulation in Healthcare, 2, 96-101. (doi: 10.1097/SIH.0b013e31805d8b0d)

King, H.B., Battles, J., Baker, D.P., Alonso, A., Salas, E., Webster, J., … Salisbury, M. (2008) TeamSTEPPS: Team strategies and tools to enhance performance and patient safety. In: K. Henrikson, J.B. Battles, M.A. Keyes, & M.L. Grady (Eds.) *Advances in patient safety: new directions and alternative approaches (vol. 3: performance and tools).* Rockville, MD: Agency for Healthcare Research and Quality, pp. 5-20.

MacDonnell, C.P., Rege, S.V., Misto, K., Dollase, P., & George, P. (2012) An introductory interprofessional exercise for healthcare students. American Journal of Pharmaceutical Education, 76, 1-6.

Maignan, M., Koch, F., Chaix, J., Phellouzat, P., Binauld, G., Muret, R.C., … Debaty, G. (2016) Team emergency assessment measure (TEAM) for the assessment of non-technical skills during resuscitation: validation of the French version. Resuscitation, 101, 115-120. (doi:10.1016/j.resuscitation.2015.11.024).

Maguire, M.B., Bremner, M.N., Yanosky, D.J. (2014) Reliability and validity testing of pilot data from the TeamSTEPPS performance observation tool. Journal of Nursing and Care, 3, 202. (doi:10.4172/2167-1168.1000202).

Malec, J.F., Torscher, L.C., Dunn, W.F., Wiegman, D.A., Arnold, J.J., Brown, D.A., … Phatak, V. (2007) The Mayo high performance teamwork scale: reliability and validity for evaluating key crew resource management skills. Simulation in Healthcare, 2, 4-10. (doi:10.1097/SIH.0b013e31802b68ee).

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D.G. (2010). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. International Journal of Surgery, 8, 336–341.

Morgan, P.J., Pittini, R., Regehr, G., Marrs, C., Haley, M.F. (2007) Evaluating teamwork in a simulated obstetric environment. Anesthesiology,106, 907-915. (doi:10.1097/01.anes.0000265149.94190.04).

Morgan, P.J., Tregunno, D., Pittini, R., Tarshis, J., Regehr, G., Desousa, S., … Milne, K. (2012) Determination of the psychometric properties of a behavioural marking system for obstetrical team training using high-fidelity simulation. BMJ Quality and Safety, 21, 78-82. (doi:10.1136/bmjqs-2011-000296).

Onwochei, D.N., Halpern, S., Balki, M. (2017) Teamwork assessment tools in obstetric emergencies- a systematic review. Simulation in Healthcare, 12, 165-176. (doi:10.1097/SIH.0000000000000210).

Oriot, D., Bridier, A., & Ghazali, D.A. (2016) Development and assessment of an evaluation tool for team clinical performance: the team average performance assessment scale (TAPAS). Health Care: Current Reviews, 4, 1000164. (doi: 10.4172/2375-4273.1000164).

Paige, J.T., Garbee, D.D., Kozmenko, V., Yu, Q., Kozmenko, L., Yang, T., … Swartz, W. (2014) Getting a head start: high-fidelity, simulation-based operating room team training of interprofessional students. Journal of the American College of Surgeons, 218, 140-149. (doi:10.1016/j.jamcollsurg.2013.09.006).

Patterson, M.D., Geis, G.L., Falcone, R.A., LeMaster, T., Wears, R.L. (2013) In situ simulation: detection of safety threats and teamwork training in a high risk emergency department. BMJ Quality and Safety, 22, 468-477. (doi:10.1136/bmjqs-2012-000942).

Passauer-Baierl, S., Hull, L., Miskovic, D., Russ, S., Sevdalis, N., & Weigl, M. (2014) Re-validating the Observational Teamwork Assessment for Surgery tool (OTAS-D): cultural adaptation, refinement, and psychometric evaluation. World J Surgery, 38, 305-13. (doi:10.1007/s00268-013-2299-8).

Phitayakorn, R., Minehart, R.D., Hemingway, M.W., Pian-Smith, M.C.M., & Petrusa, E. (2015) The relationship between intraoperative teamwork and management skills in patient care. Surgery, 158, 1434-1440. (doi:10.1016/j.surg.2015.03.031).

Rosen, M.A., Salas, E., Wilson, K.A., King, H.B., Salisbury, M., Augenstein, J.S., … Birnbach, D.J. (2008) Measuring team performance in simulation-based training: adopting best practices for healthcare. Simulation in Healthcare, 3, 33-41. (doi:10.1097/SIH.0b013e3181626276).

Rovamo, L., Nurmi, E., Mattila, M.M., Suominen, P., & Silvennoinen, M. (2015) Effect of a simulation-based workshop on multidisciplinary teamwork of newborn emergencies: an intervention study. BMC Research Notes, 8, 671. (doi:10.1186/s13104-015-1654-2).

Sigalet, E., Donnon, T., Cheng, A., Cooke, S., Robinson, T., Bissett W, … Grant, V. (2013) Development of a team performance scale to assess undergraduate health professionals. Academic Medicine, 88, 989-996. (doi:10.1097/ACM.0b013e318294fd45).

Sigalet, E.L., Donnon, T.L., & Grant, V. (2015) Insight into team competence in medical, nursing and respiratory therapy students. Journal of Interprofessional Care, 29, 62-67. (doi:10.3109/13561820.2014.940416).

Tregunno, D., Pittini, R., Haley, M., & Morgan, P.J. (2009) Development and usability of a behavioural marking system for performance assessment of obstetrical teams, Quality and Safety in Health Care, 18, 393-396. (doi:10.1136/qshc.2007.026146).

Walker, S., Brett, S., McKay, A., Lambden, S., Vincent, S., & Sevdalis, N. (2011) Observational skill-based clinical assessment tool for resuscitation (OSCAR): development and validation. Resuscitation, 82, 835-844. (doi:10.1016/j.resuscitation.2011.03.009).

Weller, J., Shulruf, B., Torrie, J., Frengley, R., Boyd, M., Paul, A., … Dzendrowskyj, P. et al. (2013) Validation of a measurement tool for self-assessment of teamwork in intensive care. British Journal of Anaesthesia, 111, 460-467. (doi:10.1093/bja/aet060).

Weller, J., Frengley, R., Torrie, J., Shulruf, B., Jolly, B., Hopley, L., … Paul, A. (2011) Evaluation of an instrument to measure teamwork in multidisciplinary critical care teams. BMJ Quality and Safety, 20, 216-222. (doi:10.1136/bmjqs.2010.041913).

Whittaker, G., Abboudi, H., Khan, M.S., Dasgupta, P., Ahmed, K. (2015) Teamwork assessment tools in modern surgical practice: a systematic review. Surgery Research and Practice, 2015, 1-11. (doi:/10.1155/2015/494827).

Zhang, C., Miller, C., Volkman, K., Meza, J., & Jones, K. (2015) Evaluation of the team performance observation tool with targeted behavioral markers in simulation-based

Interprofessional education. Journal of Interprofessional Care, 29, 202-208.

(doi:10.3109/135661820.2014.982789).

Figure 1. PRISMA search and selection pathway for included studies

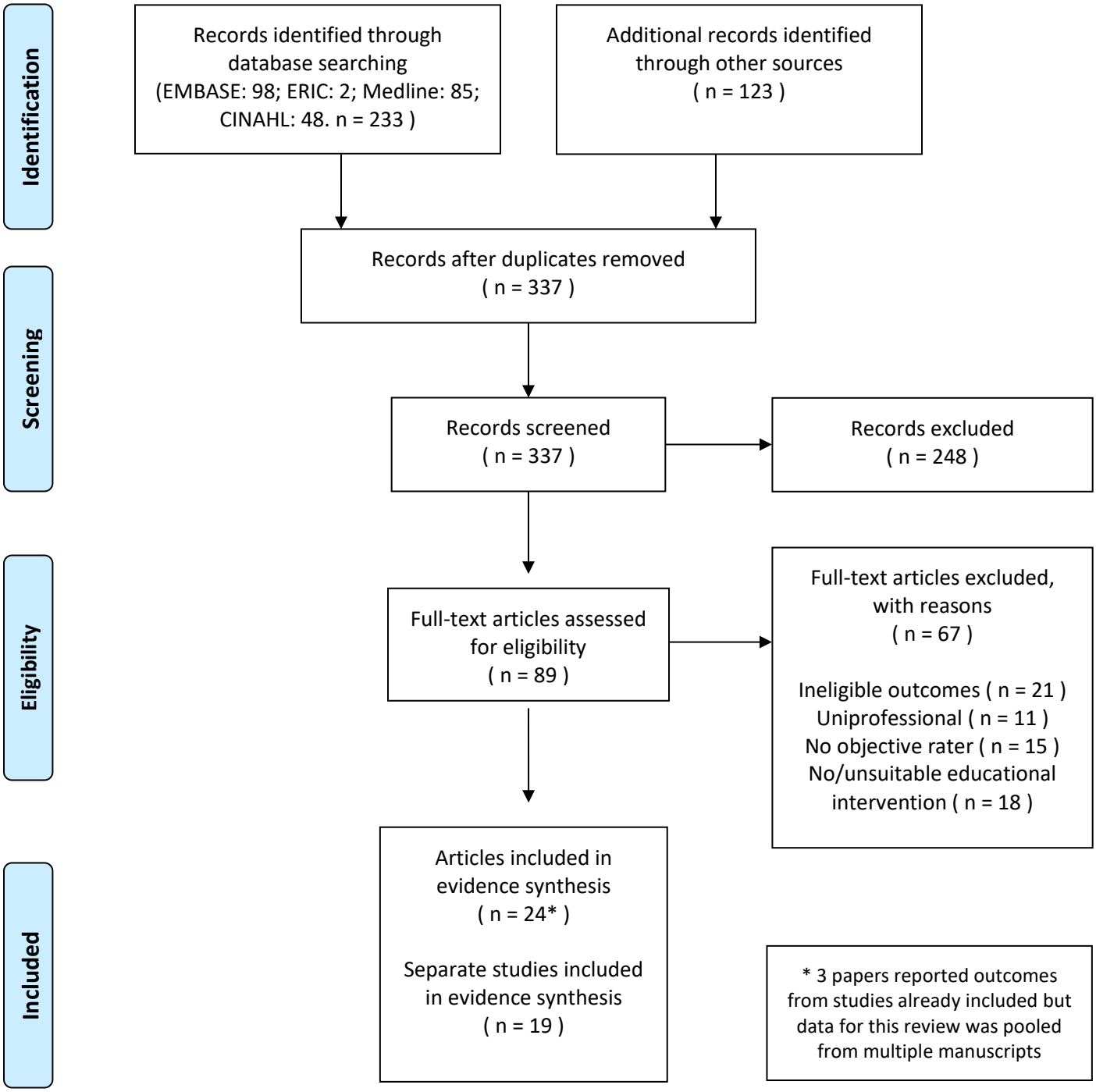

Table 1. Matrix of Included Studies

Study Details	Population Studied	Objective/Outcome measure	Raters and Rating Process	Psychometric Properties/Outcomes
Anaesthetists' Non-Technical Skills (ANTS) behavioural scale				
Patterson, M.D., et al. (2013) U.S.A.	Paediatricians, nurses, paramedics, respiratory therapists, patient care assistants, others (n = 218)	Primary outcome was the number and type of Latent Safety Threats identified during sims. Secondary measures were participants' assessment of impact on patient care, value to participants, quality of teamwork assessed with ANTS.	Rater: (n = 1) Training: Yes Rating Process: Blinded retrospective, independent video review	No significant improvement in team behaviours during study period, although mean scores noted to be high throughout. Previously validated by Fletcher, G., et al. (2003) for anaesthetists, but no further psychometric properties tested.
Jankouskas, T., et al. (2007) U.S.A.	Paediatricians, anaesthetists, nurses (n = 140 participants in 7 groups)	Primary outcome was perceived levels of collaboration and satisfaction about care decisions using a tool for this purpose. As a secondary outcome measure teamwork was measured using the ANTS.	Raters: paediatric nurse and anaesthetics resident (n=2) Training: Not stated Rating process: Blinded to scenario order, retrospective independent video review	Previous validation by Fletcher, G., et al. (2003) cited, but no further psychometric properties tested.

Assessment of Obstetrical Team Performance (AOTP) and Global Assessment of Obstetrical Team Performance (GAOTP)				
Morgan, P.J., et al. (2012) Canada	Anaesthetists, nurses, obstetricians, family doctors (n = 136 'performances' from 10 teams)	To validate a behavioural marking tool in obstetric high-fidelity simulations.	Raters: 3 nurses, 1 midwife, 2 anaesthetists, 2 obstetricians (n = 8) Training: 8 hour session Rating Process: Retrospective, independent video review	Debriefing after 2nd or 3rd sessions did not affect teamwork scores significantly (p=0.6). 1088 evaluations completed (136 performances x 8 raters). Internal consistency for AOTP was 0.96 and GAOTP was 0.91 with Cronbach's alpha. Collectively as a 22-point scale this was 0.97. Acceptable interrater reliability with 8 raters (single rater ICC 0.81)
Emergency Team Dynamics scale				
Bradley, P., et al. (2009) U.K.	Medical and nursing students (n = 30 students in the interprofessional arm)	To identify the effects on interprofessional resuscitation skills teaching on students' attitudes, leadership, team working and performance skills.	Raters: 1 author rated all videos, a 2nd author rated 10% sample (n = 2) Training: Not stated Rating Process: Retrospective, video review	Previous validation cited but not stated or accessible via references, no further psychometric properties tested.

				There was no significant difference in performance between inter- and uniprofessional groups in the ETD scores.
Human Factors Rating Scale (HFRS) and Global Rating Scale (GRS)				
Morgan, P.J., et al. (2007) Canada	Obstetricians, anaesthetists, obstetric nurses (n = 34 participated in 12 simulations)	To determine if 2 new rating scales could reliably assess obstetric team performance	Raters: Healthcare professionals with experience in obstetrics or human factors (n = 9) Training: not stated Rating Process: Retrospective, independent video review	Single rater ICC for the HFRS was low (0.341) but collectively Cronbach's alpha was 0.823.For the GRS the single rater ICC was 0.446 compared to 9-rater Cronbach's alpha of 0.879. Pearson product-moment correlation between HFRS and GRS was 0.934 suggesting they were measuring similar constructs.
KidSIM Team Performance Scale (KidSIM)				
Sigalet, E., et al. (2013) Canada	Nursing, medical and respiratory therapy students (n = 196)	To assess the impact of team training in addition to team simulation for teamwork scores in multiprofessional student teams using KidSIM team	Raters: 2 doctors, 2 nurses, 1 respiratory therapist (n = 6) Training: Not stated	Improved mean aggregate performance scores from sim 1 to 2 in both groups (paired t-tests), smaller mean effect size in

		performance scale. Does simulation training improve scores and is this heightened by the additional use of teamwork training prior to simulations?	Rating Process: Retrospective, video review, all raters reviewed all content.	intervention group (Cohen's d= 0.56 vs 0.28 with p<0.001 and p<0.05) Good internal reliability (Cronbach's Alpha=0.9) and factor analysis performed
Mayo High Performance Teamwork Scale (Mayo Scale)				
Weller, J., et al. (2011) New Zealand **"Adapted Mayo Scale"**	Doctors and nurses (n = 40 teams of 4)	To develop and validate an instrument to measure teamwork behaviours in critical care.	Raters: Anaesthetists or Critical Care clinicians (n = 3) Training: Yes Rating process: Retrospective video review	Exploratory Factor Analysis clustered items into 3 categories, all gave acceptable-good Cronbach's alpha (internal consistency). A significant improvement was seen in performance with time and seniority (implying construct validity). Sims led by specialists over trainees have a statistically significantly higher team score (p<0.001).
Burton, K.S., et al. (2011) U.S.A.	Nurses and respiratory therapists (n = 19)	To assess whether simulation would improve technical and non-technical skills in dealing with ECMO circuit emergencies	Raters: Not stated (n = 2 per simulation but not known if same raters for all)	Moderate correlation was found between reviewers (Pearson's correlation coefficient=0.41, p<0.001).

		and allow transfer skills from simulation to the clinical setting.	Training: Review of original tool publication, didactic session and group video review. Rating Process: Randomised, retrospective video review	Scores improved through the quarters (but only significantly so from 1st to 2nd quarter). Previously validated by Malec, et al. (2007)
Hobgood, C., et al. (2010) U.S.A. **"Adapted Mayo Scale"**	Medical and nursing students (n = 80 in simulation cohort)	To conduct a RCT of four pedagogical methods to deliver teamwork training and measure the effects of each method of student teamwork knowledge, skills and attitudes	Raters: Independent (n = 7) Training: Yes Rating Process: Randomised, retrospective, independent video review of a sample of all videos	The revised 20-item Mayo HPTS had inter-rater reliabilities with ICC from 0.83-1.0 on 19/20 items. There were no significant differences between cohorts with ANOVA (p=0.134) Tool was previously validated by Malec, et al. (2007)
Objective Teamwork Assessment System (OTAS)				
Phitayakorn, R., et al. (2015) U.S.A.	Nurses, scrub technicians, anaesthetic residents, surgical residents (n = 25)	To explore the correlation between operating theatre teamwork and adherence to patient care guidelines; to assess the psychometrics of a range of teamwork tools for surgical in situ simulation	Raters: 2 anaesthetists, 1 surgeon, 1 scrub nurse and a social scientist (members of the simulation team) (n = 5)	No relationship was found between technical and non-technical skill usage in the sims. High OTAS scores were given for some teams who did not complete the

			Training: 1 hour session and reviewed original paper generating tool Rating process: Retrospective video review	majority of clinical tasks on the checklist. Interrater agreement was 0.42-0.9 (mean 0.7). No further exploration of psychometrics, but tool described as externally validated (Passauer-Baierl, et al., 2014)
Observational Skill-based Clinical Assessment tool for Resuscitation (OSCAR)				
Walker, S., et al. (2011) U.K.	Anaesthetists, nurses, physicians in 8 simulations: 4 in the simulation suite and 4 in the hospital environment	To develop a feasible and psychometrically sound tool to assess team behaviours during cardiac arrest resuscitation attempts	Raters: 2 'expert clinical observers' Training: not stated Rating process: Retrospective, independent	Internal consistency was acceptable to good with Cronbach's alpha ranging from 0.736-0.965, 15/18 items had Cronbach's alpha >0.8. ICC ranged from 0.652-0.911, for individual domains and 0.767-0.807 overall ($p<0.001$) demonstrating good inter-rater reliability.

Other / Unnamed Instruments				
Auerbach, M., et al. (2014) U.S.A. **"A validated trauma simulation evaluation tool"**	Nurses, Physician Assistants, Physicians (student, resident, fellows, attendings), allied health, social workers, diagnosticians, transport (n = 398)	To evaluate the feasibility and measure the impact of an in situ interdisciplinary paediatric trauma quality improvement simulation program using a behavioural marker tool	Raters: Lead investigator who also developed scenarios, ran simulation and debriefed (n = 1) Training: Not stated Rating process: Contemporaneous	Overall performance, teamwork scores and clinical markers/checklist items improved over time (this was statistically significant). Psychometric properties not stated
Daniels, K., et al. (2008) U.S.A. **"Faculty developed Checklist of Expected Actions"**	Nurses, anaesthetic resident, obstetric residents (n = 49)	Can a simulation of obstetric crises be created for team training? Can simulation identify clinical performance deficiencies of obstetric residents that can serve as a basis for focused teaching?	Raters: Faculty Obstetricians (n = 2) Training: Yes, multiple sessions Rating process: Retrospective video review	Checklist internally but not formally validated. Learning points derived but conclusions cannot be drawn from this due to study design. Psychometric properties not stated
MacDonnell, C.P., et al. (2012) U.S.A. **"Teamwork global rating**	Medical students, nursing students, pharmacy students in mixed teams of 3 (n = 251)	To evaluate healthcare students' perceptions of an introductory interprofessional exercise and their team dynamics.	Raters: Faculty from the medical school Training: not stated Rating process: Contemporaneous	Team dynamics was rated from poor to excellent (poor 0%, fair 21%, good 36%, excellent 31%, outstanding 12%). Psychometric properties not stated

scale based on a validated evaluation instrument"				
Operating Room Team Assessment Scale (ORTAS)				
Paige, J.T., et al. (2014) U.S.A.	Medical students, nursing anaesthesia students, nursing students in teams of 6 attended 2 simulations (n = 66)	To evaluate the immediate impact of conducting interprofessional student operating room team training using high-fidelity simulation on students' team-related attitudes and behaviours.	Raters: 3-4 trained observers rated each scenario Training: 2 hour session Rating process: Contemporaneous	Acceptable relative and absolute coefficients were demonstrated for multiple raters with generalisability coefficients of 0.94-0.95 for 3-4 raters. Mean observer rating scores improved from 1st to 2nd scenario.
TAPAS				
Oriot, D., et al. (2016) France	1st team: multiprofessional teams of 4 (physicians, nurses, ambulance drivers) 2nd team (control): excluded as emergency physicians only was used to compare	To develop and psychometrically assess a clinical evaluation tool for simulated adult, neonatal and paediatric emergencies.	Raters: Doctors (1 paediatric intensivist, 3 paediatric emergency physicians, 1 anaesthetist, 3 emergency physicians) (n = 8) Training: 2 hour session Rating process: Contemporaneous, 2 raters per simulation	Acceptable internal consistency (Cronbach's alpha 0.745, from 0.646-0.806 for various items) and modest correlation coefficient 0.64. Intraclass correlation coefficient was 0.862 (high reproducibility) Internal consistency and reliability assessed by

	teamwork skills at 4 months (n = 48)			Ghazali, et al. (2016) in Sim-Stress study
Team Emergency Assessment Measure (TEAM)				
Cooper, S., et al. (2010) U.K. and Australia	Medical students and nursing students U.K.: 53 video recorded hospital simulations, number of participants not stated. Australia: Teams of 5, completing 3 simulations (n = 15)	To develop a valid, reliable and feasible teamwork assessment measure for emergency resuscitation team performance.	Raters: 2 doctors, 4 nurses/resuscitation officers with 15-26 years acute care experience (n = 6) Training: Not stated Rating process: Contemporaneous for Australia arm, retrospective for U.K. simulations, 1 rater per simulation, with 11% having second scoring by separate rater	Content validity index was acceptable at 0.83 for individual items and 0.96 overall. Acceptable construct and concurrent validity between total item score and global rating (rho 0.95, $p<0.01$) were demonstrated. Cronbach's alpha coefficient = 0.89 demonstrating high internal consistency
Rovamo, L., et al. (2015) Finland	Paediatricians, anaesthetists, obstetricians, midwives, neonatal nurses in 2 units (n = 99)	To compare the TEAM scores in simulations between a control group (sim only) and an intervention group (lecture + sim) to evaluate the impact of CRM and ANTS instruction on teamwork during simulated newborn emergencies.	Raters: Anaesthetists (n = 3) Training: Not stated Rating process: Retrospective, independent video review.	TEAM scoring has good internal consistency (Cronbach's alpha 0.919, $p<0.01$), moderate index of agreement between raters (0.41). Inter-rater reliability was poor to moderate. Tool previously validated by Cooper, et al., 2010

Team Performance Observation Tool (TPOT)				
Zhang, C., et al. (2015) U.S.A.	Physical therapy and nursing students (n = 72)	To decrease the subjectivity of the TPOT and determine its psychometrics when using scenario-specific targeted behavioural markers.	Raters: Nurses and physical therapist (manuscript authors) (n = 3) Training: Not stated, but raters trained in teamwork Rating Process: Retrospective, independent video review.	Higher TPOT overall ratings were associated with fewer team errors ($p=0.008$). The addition of scenario-specific targeted behavioural markers improves the validity and reliability of the TPOT over the scale alone. Good test re-test reliability ($k=0.707$, $p<0.001$), interrater reliability ($k=0.73$), Cronbach's alpha 0.92

Acknowledgements

Study completed as part of a Masters Dissertation in Clinical Education at Plymouth University.

Funding: none

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of this article.