A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment

Jonathan S Ilgen,[1] Irene W Y Ma,[2] Rose Hatala[3] & David A Cook[4,5]

CONTEXT The relative advantages and disadvantages of checklists and global rating scales (GRSs) have long been debated. To compare the merits of these scale types, we conducted a systematic review of the validity evidence for checklists and GRSs in the context of simulation-based assessment of health professionals.

METHODS We conducted a systematic review of multiple databases including MEDLINE, EMBASE and Scopus to February 2013. We selected studies that used both a GRS and checklist in the simulation-based assessment of health professionals. Reviewers working in duplicate evaluated five domains of validity evidence, including correlation between scales and reliability. We collected information about raters, instrument characteristics, assessment context, and task. We pooled reliability and correlation coefficients using random-effects meta-analysis.

RESULTS We found 45 studies that used a checklist and GRS in simulation-based assessment. All studies included physicians or physicians in training; one study also included nurse anaesthetists. Topics of assessment included open and laparoscopic surgery (n = 22), endoscopy (n = 8), resuscitation (n = 7) and anaesthesiology (n = 4). The pooled GRS–checklist correlation was 0.76 (95% confidence interval [CI] 0.69–0.81, n = 16 studies). Inter-rater reliability was similar between scales (GRS 0.78, 95% CI 0.71–0.83, n = 23; checklist 0.81, 95% CI 0.75–0.85, n = 21), whereas GRS inter-item reliabilities (0.92, 95% CI 0.84–0.95, n = 6) and inter-station reliabilities (0.80, 95% CI 0.73–0.85, n = 10) were higher than those for checklists (0.66, 95% CI 0–0.84, n = 4 and 0.69, 95% CI 0.56–0.77, n = 10, respectively). Content evidence for GRSs usually referenced previously reported instruments (n = 33), whereas content evidence for checklists usually described expert consensus (n = 26). Checklists and GRSs usually had similar evidence for relations to other variables.

CONCLUSIONS Checklist inter-rater reliability and trainee discrimination were more favourable than suggested in earlier work, but each task requires a separate checklist. Compared with the checklist, the GRS has higher average inter-item and inter-station reliability, can be used across multiple tasks, and may better capture nuanced elements of expertise.

Medical Education 2015: 49: 161–173
doi: 10.1111/medu.12621

Discuss ideas arising from the article at "www.mededuc.com discuss"

[1]Division of Emergency Medicine, Department of Medicine, University of Washington School of Medicine, Seattle, Washington, USA
[2]Department of Medicine, University of Calgary, Calgary, Alberta, Canada
[3]Department of Medicine, University of British Columbia, Vancouver, British Columbia, Canada
[4]Mayo Clinic Multidisciplinary Simulation Center, Mayo Clinic College of Medicine, Rochester, Minnesota, USA
[5]Division of General Internal Medicine, Mayo Clinic, Rochester, Minnesota, USA

Correspondence: Jonathan S Ilgen, Division of Emergency Medicine, University of Washington School of Medicine, Harborview Medical Center, 325 9th Avenue, Box 359702, Seattle, Washington 98104-2499, USA. E-mail: ilgen@u.washington.edu

INTRODUCTION

Checklists and global rating scales (GRSs) are frequently used in assessment in health professional education, and the relative advantages and disadvantages of these two types of tool have long been debated.[1–4] Checklists prompt raters to attest to the performance or omission of directly observable actions, whereas GRSs typically asks raters to judge participants' overall performance or to provide global impressions of performance on sub-tasks. Checklists are relatively intuitive to use and – especially for raters who are less familiar with the clinical task at hand – provide step-by-step outlines for observable behaviours and guidance for formative feedback.[5] Although checklists offer the allure of a more 'objective' frame of measurement, evidence suggests that this format may not necessarily confer greater validity or reliability.[6,7] By requiring raters to dichotomise ratings, checklists may result in a loss of information,[1,8] and this format may reward thoroughness at the expense of actions that more accurately reflect clinical competence.[6,9,10] By contrast, GRSs have been shown to detect differing levels of expertise more sensitively than the checklist,[11] although the rendering of accurate global impressions requires subjective rater judgement and decision making.[12] While this subjectivity is likely to have value,[13,14] the reliability and accuracy of assessments may be dependent upon rater characteristics, such as familiarity with the scale, clinical expertise, training and personal idiosyncrasies, and on the complexity of the task, which leads some to question the defensibility of expert global impressions in high-stakes assessment settings.[15–17]

Two seminal reviews conducted 23 years ago considered the strengths and weaknesses of checklists and GRSs.[6,10] The validity evidence and psychometric properties of checklists and GRSs have since been reviewed in the contexts of the objective structured clinical examination (OSCE)[18–20] and direct clinical observation.[17] Prior reviews, however, have not aimed to *systematically* compare the validity evidence supporting the interpretations of these scales' scores. A more systematic approach to study identification would permit comprehensive comparisons of the validity evidence supporting the interpretations of scores on these scales. An updated review might also incorporate recently published studies, distinguish among different facets of reliability (e.g. raters, items and stations), and allow a meta-analytic summary of quantitative data. Educators would benefit from a clearer understanding of GRSs' and checklists' development and implementation processes, of their psychometric performance across studies, and of validity evidence supporting their use. This information would enable them to develop assessments in light of recent evidence, trust and defend their assessment decisions, and tailor their assessments to the needs of their learners. We recently completed a systematic review of simulation-based assessment tools,[21] which offered the opportunity to address this gap. Thus, we conducted a systematic review and meta-analysis of validity evidence for checklists and GRSs used to assess health professionals in the context of simulation-based medical education.

Research questions

The studies included herein are a subset of those reported in a previous review,[21,22] but we collected new, detailed data to answer the current research questions.

1 What are the inter-rater, inter-item and inter-station reliabilities of global ratings in comparison with checklist scores?
2 How well do global ratings and checklist scores correlate?
3 What validity evidence has been reported for global ratings and checklist scores?

The prevailing view has held that the GRS offers greater reliability than the checklist.[6,7,10] In the present review, we sought evidence to confirm or refute this opinion.

METHODS

We planned and conducted this review in adherence to the PRISMA (*p*referred *r*eporting *i*tems for *s*ystematic reviews and *m*eta-*a*nalyses) standards of quality for reporting systematic reviews.[23]

Study eligibility

From the studies included in an earlier systematic review of simulation-based assessment,[21,22] we identified original research studies published in any language that evaluated both a GRS and a checklist while using technology-enhanced simulation to assess health professions learners. In line with our prior review,[24] we defined technology-enhanced simulation as an: 'educational tool or device with which

the learner physically interacts to mimic an aspect of clinical care for the purpose of teaching or assessment. This includes (but is not limited to) high-fidelity and low-fidelity manikins, part-task trainers, virtual reality (including any computer simulation that requires non-standard computer equipment), animal models, and human cadaveric models used for teaching purposes.'[24] The use of human standardised patients was not included in this definition, although the inter-rater reliability of GRSs and checklists with standardised patients in the OSCE setting was recently reported.[18] We defined checklists as instruments with a dichotomous response format and more than one item; we excluded studies with only a single checklist item (i.e. an overall pass/fail only). We defined GRSs as instruments with more than two response options per item. Because these scales have been designed to allow global judgements, we included single-item summative GRSs (i.e. those that ask for a 'global impression'). To enable meaningful psychometric comparisons between instrument types, we excluded studies in which the GRS and checklist assessed different constructs.

Study identification and selection

Our search strategy has been previously published in full.[22,24] To summarise briefly, an experienced research librarian developed a search strategy that included terms focused on the topic (e.g. simulat*), learner population (med*, nurs*, health occupations), and assessment (e.g. assess*, valid*). We used no beginning date cut-off, and the last date of search was 26 February 2013. Reviewers worked independently in pairs to screen studies for inclusion, resolving conflicts by consensus. From this large pool of studies, we identified studies that included both a GRS and a checklist that were designed to assess the same construct (reviewer inter-rater reliability [IRR] intraclass correlation coefficient [ICC]: 0.76). Some articles referred to more than a single checklist or a single GRS; in these instances, we selected tools using a hierarchy described previously.[21]

Data extraction

We abstracted data independently and in duplicate for all variables, resolving conflicts by consensus. Inter-rater reliability was substantial for nearly all variables (Appendix S1 [online] gives details on IRR for data abstraction). We noted the task being assessed, and classified the assessment as measuring technical skills (the learner's ability to demonstrate a procedure or technique) or non-technical skills (such as communication or team leadership). We collected validity evidence from five sources[25] using previously defined operational definitions,[22,26] which included: internal structure (inter-rater, inter-item and inter-station reliability, and factor analysis); content (processes used to ensure that items accurately represent the construct); relations to other variables (association with participant characteristics, such as training level, or scores from another instrument); response process (the relationship between the intended construct and the thought processes of participants or observers), and consequences (the assessment's impact on participants and programmes).[26] We also coded information about the raters, instruments and tasks, and evaluated study quality using the Medical Education Research Study Quality Instrument[27] (MERSQI) (IRR ICCs: 0.51–0.84[21]).

Data analysis

We pooled reliability and correlation coefficients using random-effects meta-analysis. We used *Z*-transformed coefficients for these analyses, and then transformed the pooled results back to the native format for reporting. We used the Spearman–Brown formula to adjust all reliability coefficients to reflect a single item, station or rater prior to pooling. We performed analyses separately for GRS and checklist scores. We quantified between-study inconsistency using the I^2 statistic; I^2 values of > 50%, 25–49% and < 25% indicate large, moderate and little inconsistency, respectively.[28] We conducted planned sensitivity analyses excluding studies with single-item GRSs. To avoid undue influence from any single instrument, we conducted *post hoc* sensitivity analyses excluding all studies using the objective structured assessment of technical skill (OSATS) GRS[29] because this tool comprised nearly one-third of our GRS sample.

Nearly all inter-rater reliability results in our sample comprised kappa, weighted kappa or an ICC corresponding to the Shrout and Fleiss (2, 1) ICC, all of which reflect the reliability of a single rater.[30] Thus, we report (2, 1) inter-rater reliability coefficients and classify these using criteria proposed by Landis and Koch[31] (fair [ICC values: 0.21–0.4], moderate [ICC values: 0.41–0.6], substantial [ICC values: 0.61–0.8]). For inter-item and inter-station reliability, which always used a (2, k) ICC (e.g. Cronbach's alpha),[30] we used the Spearman–Brown formula to adjust pooled coefficients to reflect the median (or near-median) number of observations, and classified

these as suboptimal (values < 0.70), good (0.70–0.89) or substantial (> 0.90).[32] We also noted instances in which authors did not clearly identify analyses as reflecting inter-item or inter-station reliability, and performed sensitivity analyses excluding these studies. We classified correlation coefficients as small (0.10–0.29), moderate (0.30–0.49) and large (≥ 0.50) using thresholds proposed by Cohen.[33] We used SAS Version 9.3 (SAS Institute, Inc., Cary, NC, USA) to perform all analyses.

RESULTS

Trial flow is shown in Appendix S2 (online). From 11 628 potentially relevant articles, we identified 45 that used a GRS and a checklist to measure the same construct, reflecting data from 1819 trainees (median: 27 trainees per study; interquartile range [IQR]: 20–55). Table 1 summarises the key features of the included studies.

All studies involved the assessment of physicians at some stage of training, primarily represented by postgraduate physician trainees (36 studies, 1188 trainees) and medical students (seven studies, 306 trainees). One study also enrolled nurse anaesthetist students[34] and another included 'industry representatives' along with trainees.[35] Thirteen studies enrolled different cohorts of trainees concurrently (e.g. medical students and postgraduates), and nearly all studies enrolled trainees across multiple postgraduate years. All 45 studies used GRSs and checklists to assess skills in the simulation setting, and one assessed patient-based outcomes as well.[36] The average MERSQI score was 13.3 (maximum on scale: 18.0; range: 10.5–16.0). (Appendix S3 shows detailed MERSQI codes.)

Scale characteristics

The clinical areas of assessment included open (n = 18) and minimally invasive (n = 5) surgery, endoscopy (n = 8), resuscitation (n = 7), anaesthesiology (n = 4), and non-technical skills for both resuscitation and surgery (n = 3) (Table 1). About two-thirds of the reports (GRS, n = 27; checklist, n = 29) included examples of their scales or provided sufficient description to allow their replication. Among studies in which item numbers were reported, GRSs (n = 43) contained an average of six items (median: seven; range: 1–13), and checklists (n = 35) contained an average of 19 items (median: 17; range: 3–49). Forty studies provided descriptions of GRS anchors, which were most commonly behavioural (i.e. directly observable actions, n = 23); other anchors included proficiency (i.e. ranging from 'high' to 'low' performance without outlining specific behaviours, n = 10), Likert scale-based anchors (i.e. ranging from 'disagree' to 'agree', n = 5), expert/intermediate/novice performance (n = 1), and visual analogue scales (n = 3) (some studies used multiple anchor types). Thirteen studies used the OSATS GRS[29] or very slight modifications of it, and another 14 studies used the OSATS as the starting point for a new instrument.

In 20 studies, the assessment was comprised of more than one station, each reflecting a unique task; among these, the median number of stations was five (range: 2–10). In 17 of these 20 studies, authors used the same GRS across all stations; two studies used a unique GRS at each station, and in one study the number of unique GRSs was unclear. By contrast, 17 of the 20 multi-station studies described unique checklists for each station; in the remaining three studies, the number of checklists was unclear.

Rater characteristics and procedures

Raters in the included studies were typically physicians (n = 34). Five studies employed other medical professionals (such as nurses, emergency medical technicians and respiratory therapists), and 11 studies did not clearly describe the backgrounds of raters. Authors typically justified their rater selection by describing these individuals' expertise in the clinical area being assessed.

Fewer than half of the included studies described rater training for the scale under study (GRS, n = 21; checklist, n = 22), and few provided evidence of rater training outcomes (GRS, n = 2; checklist, n = 1). Five studies provided different degrees of rater training for GRSs than for checklists. Among the studies in which no specific training in the tool under study was reported (GRS, n = 24; checklist, n = 23), a few reported that their raters were 'experienced' (GRS, n = 4; checklist, n = 3) without further explanation of training experience.

The GRS and checklist were completed by the same rater in 39 of the 45 studies. About half of the ratings were performed live (GRS, n = 22; checklist, n = 22), and the remaining ratings were performed retrospectively using video (GRS, n = 24; checklist, n = 24); one study used both live and video reviews.[37] Raters assessed all trainees in two-thirds of the studies (n = 29); among studies in which raters

Table 1 Details of study characteristics, validity evidence and rater characteristics

				Validity evidence								
				Global rating scales‡			Checklists‡			Rater characteristics§		
Study	Participants, n, type*	Clinical task	Study quality†	IS	RoV	C, RP, CQ	IS	RoV	C, RP, CQ	Selection	Training	Blinding
Jansen et al. (1997)[46]	71, MDs	Resuscitation	15.5	R	M		R	M	C	O	G, C	
Martin et al. (1997)[29]	20, PGs	Surgery, open	14.0	R, S, O	M, T	C	R, S, O	M, T	C	E	G, C	R
Reznick et al. (1997)[47]	48, PGs	Surgery, open	13.5	S	T	C	S	T	C	E	NT	
Regehr et al. (1998)[2]	53, PGs	Surgery, open	13.0	S	M, T	C	S	M, T	C	E	G, C	R
Friedlich et al. (2001)[48]	47, PG	Surgery, open	13.5	R, S	M, T	C	R, S	M, T	C	E	NT	R
Morgan et al. (2001)[49]	145, MSs	Anaesthesia	11.5	R	M		R	M	C	O	G, C	R
Murray et al. (2002)[50]	64, MSs, PGs	Resuscitation	13.5	R	M, T	C	R, O	M, T	C	E	G,	R
Adrales et al. (2003)[51]	27, MSs, PGs, MDs	Surgery, MIS	12.5		T	C		T		E	G, C	R, T
Datta et al. (2004)[52]	56, PGs, MDs	Surgery, open	13.0	R	M, T	C	R	M, T	C	NR	G, C	R, T
Murray et al. (2004)[53]	28, PGs	Anaesthesia	13.5	R, S	M, T		R, S	M, T	C	E	G, C	R
Weller et al. (2004)[54]	71, MSs	Resuscitation + NTS	12.0		T			T		E	NT	R
Bann et al. (2005)[55]	11, PGs	Surgery, open + MIS	14.0	R, O	T	C	R, O	T	C	NR	NT	R, T
Moorthy et al. (2005)[56]	27, PGs	NTS (surgery)	13.5	R	M, T	C		T	C	E	NT	R
Murray et al. (2005)[34]	43, PGs, NAs	Anaesthesia	14.5	R, S	T	C	R, S	T	C	E	G, C	R
Berkenstadt et al. (2006)[57]	145, PGs	Anaesthesia	12.5	R, S	M	CQ	R, S	M	C, CQ	NR	NT	R
Broe et al. (2006)[58]	20, PGs	Surgery, MIS	13.5	R	T	C	R	T	C	E	NT	R, T
Matsumoto et al. (2006)[59]	16, PGs	Endoscopy	13.0	O	M, T	C		M, T	C	NR	NT	T
Banks et al. (2007)[60]	20, PGs	Surgery, MIS	12.5	R, O	M, T	C	R, O	M, T	C	E	G, C	

Table 1 (Continued)

				Validity evidence								
				Global rating scales‡			Checklists‡			Rater characteristics§		
Study	**Participants, n, type***	**Clinical task**	**Study quality†**	**IS**	**RoV**	**C, RP, CQ**	**IS**	**RoV**	**C, RP, CQ**	**Selection**	**Training**	**Blinding**
Fialkow et al. *(2007)*[61]	55, PGs	Endoscopy	14.0	R, I	T	C	R, I	T	C	E	C	
Goff et al. *(2007)*[62]	13, PGs, MDs	Endoscopy	13.0	R, I	M, T	C	R, I	M, T	C	O	G, C	
Khan et al. *(2007)*[63]	65, PGs, MDs	Surgery, open	13.5		M, T	C		M, T	C	E	NT	R
Zirkle et al. *(2007)*[64]	19, PGs	Surgery, open	13.5	R	M, T	C	R	M, T	C	E	NT	R, T
Leung et al. *(2008)*[65]	16, PGs, MDs	Endoscopy	15.5	R	M, T		R	M, T		E	NT	R, T
Siddiqui et al. *(2008)*[66]	40, PGs	Surgery, open	14.0	R, O	M, T	C	R, O	M, T	C	E	NT	R, T
Chipman & Schmidz (2009)[67]	25, PGs	Surgery, open	13.0	I, O	T	C	I, O	T	C	NR	G, C	
Huang et al. *(2009)*[68]	42, PGs	Venous access	12.5		M			M	C, RP, CQ	O	G, C	
Insel et al. *(2009)*[69]	68, PGs, MDs	Surgery, MIS	12.5		T	C		T	C	R	G, C	T
LeBlanc et al. *(2009)*[70]	32, MSs, PGs	Surgery, open	13.5	R, S	M, T	C	R, S	M, T		E	NT	R
White et al. *(2010)*[35]	20, unclear	Endoscopy	10.5		T	C		T	C	E	NT	
Gordon et al. *(2010)*[37]	17, PGs	Resuscitation	12.5		T	C			C	R	G, C	R
Faulkner et al. *(1996)*[71]	12, PGs	Surgery, open	13.0		M	C		M	C	E	NT	R
Siddighi et al. *(2007)*[72]	26, PGs	Surgery, open	14.5	R, I	M, T	C	R, I	M, T	C	NR	G, C	
Adler et al. *(2011)*[73]	77, PGs	Resuscitation	16.0	R, S	T	C	R, S	T	C	E	G, C	R
Tuchschmid et al. *(2010)*[74]	6, unclear	Endoscopy	13.5	I	M	C	R, I	M	C	E	NT	T
Ault et al. *(2001)*[75]	77, PGs	Surgery, open	12.5	S	T	C	S	T	C	E	NT	
Khan et al. *(2003)*[76]	93, MSs, PGs, MDs	Surgery, open	11.5		M, T	C		M, T	C	E	NT	R

Table 1 (Continued)

				Validity evidence								
				Global rating scales‡			Checklists‡			Rater characteristics§		
Study	Participants, n, type*	Clinical task	Study quality†	IS	RoV	C, RP, CQ	IS	RoV	C, RP, CQ	Selection	Training	Blinding
Ponton-Carss et al. *(2011)*[77]	14, PGs	Surgery, open + NTS	13.5		M, T	C	I	M, T	C	E	G, C	
Finan et al. *(2012)*[36]	13, PGs	Airway	14.0			C	O		C	R	G, C	
Fleming et al. *(2012)*[78]	15, PGs, MDs	Endoscopy	13.5	R	M, T	C	R	M, T	C	O	NT	R, T
Hall et al. *(2012)*[79]	21, PG	Resuscitation	13.5	R	M, T	C	R	M, T	C	E	G, C	R
Jabbour et al. *(2012)*[80]	23, MSs, PGs, MDs	Endoscopy	13.5	R	T	C	R	T	C	E	NT	R, T
Ma et al. *(2012)*[81]	34, PGs	Venous access	13.5	R, I, O	M	C, RP	R, I, O	M	C, RP	R	G, C	R, T
Nimmons et al. *(2012)*[82]	20, PGs, MDs	Surgery, open	13.5	R	T	C	R	T	C	E	NT	R
VanHeest et al. *(2012)*[83]	27, PGs	Surgery, open	13.5	S	M, T	C	S	M, T	C	E	NT	
Cicero et al. *(2013)*[84]	37, PGs	Resuscitation	13.5	R, O	T	C	R, O	T	C	E	C	R

*MIS = minimally invasive surgery; NTS = non-technical skills; MDs = practising physicians; PGs = postgraduate resident physicians; MSs = medical students; NAs = nurse anaesthetists.
†Study quality evaluated using the Medical Education Research Study Quality Instrument (MERSQI); maximum score 18 (see Appendix S3 for detailed results).
‡IS = internal structure (R = inter-rater reliability; I = inter-item reliability; S = inter-station reliability; O = other); RoV = relationship to other variables (M = another measure; T = trainee characteristic); C = content; RP = response process; CQ = consequences.
§Rater characteristics: E = expert; R = researcher; O = other; NR = not reported; G = raters trained to use GRS; C = raters trained to use checklist; NT = not trained; Blinding: R = blinded to other raters; T = blinded to trainee characteristics.

assessed a subset of trainees, assignment was randomised in three studies, and was either non-random or not reported in 13 studies.

Ratings were performed in duplicate for all trainees in about two-thirds of the studies (GRS, $n = 28$; checklist, $n = 28$), and three studies performed duplicate ratings on a subset of trainees. In these 31 studies with duplicate ratings, the vast majority of raters were blinded to one another's scores (GRS, $n = 29$; checklist, $n = 28$). It was less commonly reported that raters were blinded to trainees' characteristics (GRS, $n = 13$; checklist, $n = 13$).

Correlation between instruments

Figure S1 (online) summarises the meta-analysis of correlation coefficients between GRSs and checklists in the 16 studies in which these analyses were available. The pooled correlation was moderate ($r = 0.76$, 95% confidence interval [CI] 0.69–0.81), with large inconsistency between studies ($I^2 = 71\%$).

Reliability evidence

Most studies (Table 1) provided some form of reliability (GRS, $n = 33$; checklist, $n = 33$), but only eight studies used generalisability analyses to evaluate reproducibility. Inter-rater reliability was reported in 27 GRS and 27 checklist studies. Several studies (GRS, $n = 6$; checklist, $n = 5$) used Cronbach's alpha to calculate IRR; we adjusted these to a single rater before analysis. Pooled analyses (Fig. S2, online) demonstrated substantial mean inter-rater reliabilities and high inconsistency for both GRSs (pooled IRR 0.78, 95% CI 0.71–0.83; $I^2 = 78\%$) and checklists (pooled IRR 0.81, 95% CI 0.75–0.85; $I^2 = 74\%$).

Inter-item reliability was reported infrequently (GRS, $n = 6$; checklist, $n = 7$) (Fig. S3, online). We excluded three checklist studies from meta-analyses because the authors did not specify the number of items or because the reliability analysis included non-dichotomous items. When pooled, GRSs demonstrated substantial inter-item reliability (0.92, 95% CI 0.84–0.95), which was higher than the inter-item reliability for checklists (0.66, 95% CI 0–0.84). There was moderate inconsistency among the GRS studies ($I^2 = 47\%$), whereas checklist results were quite similar ($I^2 = 0\%$).

Ten studies reported inter-station reliability (Fig. S4, online), with a median of six stations per assessment (range: 3–8). Tasks differed across stations in all 10 studies; nine of these studies used the same GRS across stations, whereas eight used a different task-specific checklist for each station (it was unclear whether the checklist differed in two studies). Pooled inter-station reliabilities were good for GRSs (0.80, 95% CI 0.73–0.85) but suboptimal for checklists (0.69, 95% CI 0.56–0.77); inconsistency was small for both scale types (GRS, $I^2 = 0\%$; checklist, $I^2 = 0\%$).

Sensitivity analyses

We conducted sensitivity analyses in several settings in which we felt that particular scale or study characteristics might bias our findings. Firstly, to ensure that the OSATS GRS (which was used in nearly a third of the studies) did not dominate our results, we conducted post hoc sensitivity analyses excluding the 13 OSATS studies. Secondly, to ensure that multi-item and single-item GRSs had similar performance characteristics, we performed sensitivity analyses excluding studies with a single-item GRS. Thirdly, to address the concern that studies with more stations and with novel checklists for each station would reduce the reliability data for checklists, we conducted sensitivity analyses limited to studies with three or more stations. Lastly, in several reports, authors did not clearly state whether analyses reflected inter-item or inter-station reliability (GRS, $n = 3$; checklist, $n = 4$). Contextual clues supported provisional classifications sufficient for the meta-analysis described above, but we also conducted sensitivity analyses excluding the ambiguous studies. For all sensitivity analyses, the results were similar to the main analyses (data not shown).

Other validity evidence

Table 1 summarises the remaining validity evidence for the included studies. Most articles provided evidence of content validity (GRS, $n = 38$; checklist, $n = 41$); for GRSs, this most commonly appeared in the form of previously reported instruments ($n = 18$), modifications of previously published instruments ($n = 15$), or expert consensus ($n = 8$), whereas for checklists, consensus among experts ($n = 26$) and modifications of prior instruments ($n = 16$) were most commonly cited.

Evidence of relations to other variables was reported in all but one study. Authors most often reported discrimination by level of training (GRS, $n = 37$; checklist, $n = 36$). As Appendix S4 shows, the checklist and GRS typically demonstrated similar discrimination by level of training, although in seven studies the GRS was more sensitive to expertise than the checklist. We found only two studies in which the checklist discriminated better than the GRS. The other source of evidence of relations to other variables was comparison with another outcome measure (GRS, $n = 28$; checklist, $n = 27$) such as 'pass/fail' judgements by raters, procedural time, hand motion analyses, and ratings of proficiency with live patients. When compared with checklists, GRSs had equivalent ($n = 11$) or higher ($n = 6$) levels of correlation to this separate measure in most studies; we found only two studies in which the checklist had the higher correlation.

Beyond reliability evidence, we found evidence of internal structure in the form of item analyses ($n = 3$), test–retest reliability ($n = 3$) and factor analysis ($n = 1$). Evidence of response process and consequences was rare or absent (for each: GRS, $n = 1$; checklist, $n = 2$).

DISCUSSION

We found moderate correlations between GRS and checklist scores, explaining on average 58% of the variance. Inter-rater reliabilities for both scale types were similarly high, whereas inter-item and inter-station reliabilities favoured the GRS. Content validity evidence was reported commonly but differed between the two scales, with GRSs referencing prior studies and checklists invoking expert opinion. Evidence for relations to other variables was usually similar for both scales, less often favoured GRSs, and rarely favoured checklists. Evidence for response process or consequences was lacking for both scales. A minority of studies reported rater training and very few provided training outcomes.

Integration with prior work

The inter-rater reliabilities for checklists were higher than those found in past investigations[38] and challenge past generalisations that checklists offer 'the illusion of objectivity. . . with very little reliability'.[7] It is conceivable that our systematic approach and large sample size permitted analyses more robust than those previously possible. Alternative explanations for these high inter-rater reliabilities include: (i) technical skills may lend themselves to more reproducible measurements than less well-defined competencies such as communication;[38] (ii) physician raters may have shared a common view of performance targets, and (iii) heterogeneity among study participants who were deliberately selected to represent different training levels may lead to artefactually high overall reliability attributable to a wider range of performance variation that was easier for raters to identify.[39] Authors did not report psychometric data stratified by training level with sufficient frequency to permit exploration of this sampling issue.

Of note, we found these high inter-rater reliabilities for both scale types despite an apparent paucity of instrument-specific rater training, contradicting, in part, literature advocating the importance of rater training.[16,40–42] These findings, together with the mixed results of earlier research on the impact of rater training,[16,42] highlight the need to further study the tasks, instruments and contexts for which rater training is needed and efficacious.

Our findings for inter-item reliability parallel those of a recent review of OSCEs, whereas inter-station reliability in that review was similar for checklists but lower for GRSs.[18] This divergence merits further exploration. We noted in our study that many checklists assessed multiple domains of competence, which may contribute to lower inter-item reliability. We suspect the low inter-station checklist reliability in our study results, at least in part, from the use of unique task-specific instruments at each station.

Early studies examining simulation-based technical skill assessment using the OSATS found better expert–novice discrimination for the GRS,[29] suggesting that judgements of expertise require more nuance than can be captured by a checklist.[1,7] Our data provide additional granularity to this interpretation, in that the two scales show similar discrimination by trainee level most of the time, yet, if one rating scale is superior, it is typically the GRS. Analyses exploring associations with other outcome measures show a similar pattern. Our finding that instrument development varied substantially between GRSs and checklists has further implications for the interpretations that can be drawn from their scores, as elaborated below.

Limitations and strengths

Our findings are tempered by limitations in both the original studies and our review methodology. The assessments in these studies represent diverse clinical topics, and task-specific checklists varied across stations and among studies. By contrast, nearly all multi-station studies used the same GRS at each station, and increased familiarity with a particular scale might favourably influence its internal consistency. As the same raters completed both scales in most studies, and the order of instrument was not consistently reported, we were unable to estimate either the direction or the magnitude of the influence of one scale rating over another. Because nearly all studies focused on technical tasks, our findings may not apply to cognitive and non-technical tasks. Our data provide insufficient granularity to explore issues of assessing specific technical skills or the influences of specific assessment conditions; these issues warrant further investigation. Finally, as stated above, within-group heterogeneity may have inflated reliability estimates.

We made a number of difficult decisions in establishing our inclusion and exclusion criteria, including our definition of technology-enhanced simulation,[24] the inclusion of diverse tasks, and the exclusion of single-item checklists. The inclusion of different articles might have altered our study conclusions. Although our inter-rater agreement was high for

most abstracted data, agreement was poor for some items as a result, at least in part, of incomplete or unclear reporting. We addressed this by reaching consensus on all reported data. The strengths of this review include its use of a broad search strategy that did not exclude potential material on the basis of language or publication year, duplicate and independent data abstraction, rigorous coding of methodological quality, and the use of reproducible inclusion criteria encompassing a broad range of learners, outcomes and study designs.

Implications for research

We found numerous instances in which authors were vague in their reporting (such as uncertainty between inter-station versus inter-item reliability) or used non-standard methods (such as in the use of Cronbach's alpha to calculate inter-rater reliability). To facilitate useful interpretations and cross-study comparisons, we encourage authors to clearly define the facet(s) of variation (raters, items, stations, time), use reliability analyses appropriate to each facet, and then explicitly report these findings. Generalisability studies may be helpful in this regard.[43]

Although our findings are generally supportive of GRSs and checklists, they clearly do not generalise to all GRSs or checklists. Yet we found several instances, particularly for the OSATS, in which authors cited evidence for a previously reported checklist in order to support their newly developed checklist. Such practices are not defensible. We remind researchers and educators that every new GRS and checklist must be validated independently and, further, that validity evidence must be collected afresh for each new application (e.g. task or learner group).[26]

Implications for practice

Our data support a more favourable view of checklists than has been suggested in earlier work.[6] Average inter-rater reliability was high and slightly better for checklists than for GRSs, and discrimination and correlation with other measures were usually similar. The use of checklists may also diminish rater training requirements and improve the quality of feedback,[41,44] although these issues require further study. However, each task requires a separate checklist and each task-specific checklist requires independent validation, especially in the context of assessing technical skills. As such, checklists will typically lag behind GRSs in the robustness of validity evidence. It is also important to highlight that, despite the perception that checklists offer more objective assessment, the construction of these tools often requires subjective judgements.

Global rating scales have important advantages. Compared with checklists, GRSs have higher average inter-item and inter-station reliability. Moreover, GRSs can be used across multiple tasks, obviating the need for task-specific instrument development and simplifying application-specific validation. Global rating scales may require more rater training, although subjective responses can capture nuanced elements of expertise[7] or potentially dangerous deviations from desired practice,[45] and reflect multiple complementary perspectives.[14] Finally, we note the inseparable interaction between the person using the instrument and the instrument itself: neither the checklist nor the GRS will supplant the need for human expertise and judgement.

Contributors: JSI contributed to the conception and design of the work, and to data acquisition and interpretation, and drafted the paper. IWYM contributed to the conception and design of the work, and to data acquisition and interpretation, and assisted in the drafting of the paper. RH contributed to the conception and design of the work, and to data acquisition and interpretation. DAC contributed to the conception and design of the work, and to data analysis and interpretation, and assisted in the drafting of the paper. All authors contributed to the critical revision of the paper and approved the final manuscript for publication. All authors have agreed to be accountable for all aspects of the work.
Acknowledgements: we acknowledge Jason Szostek, and Amy Wang (Division of General Internal Medicine), Benjamin Zendejas (Department of Surgery), and Patricia Erwin (Mayo Libraries), Mayo Clinic College of Medicine, Rochester, MN, USA; and Stanley Hamstra, Academy for Innovation in Medical Education, Faculty of Medicine, University of Ottawa, Ontario, Canada for their assistance in the literature search and initial data acquisition. We thank Glenn Regehr, Associate Director of Research and Senior scientist, Centre for Health Education Scholarship, Faculty of Medicine, University of British Columbia, Vancouver, Canada, for his constructive critique and insights.
Funding: this work was supported by intramural funds, including an award from the Division of General Internal Medicine, Mayo Clinic.
Conflicts of interest: none.
Ethical approval: not applicable.

REFERENCES

1 Hodges B, Regehr G, McNaughton N, Tiberius R, Hanson M. OSCE checklists do not capture increasing levels of expertise. *Acad Med* 1999;**74**:1129–34.

2 Regehr G, MacRae H, Reznick R, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Acad Med* 1998;**73**:993–7.
3 Ringsted C, Østergaard D, Ravn L, Pedersen JA, Berlac PA, van der Vleuten CP. A feasibility study comparing checklists and global rating forms to assess resident performance in clinical skills. *Med Teach* 2003;**25**:654–8.
4 Swanson DB, van der Vleuten CP. Assessment of clinical skills with standardised patients: state of the art revisited. *Teach Learn Med* 2013;**25** (Suppl 1):17–25.
5 Archer JC. State of the science in health professional education: effective feedback. *Med Educ* 2010;**44**:101–8.
6 van der Vleuten CPM, Norman GR, De Graaff E. Pitfalls in the pursuit of objectivity: issues of reliability. *Med Educ* 1991;**25**:110–8.
7 Norman G. Checklists vs. ratings, the illusion of objectivity, the demise of skills and the debasement of evidence. *Adv Health Sci Educ Theory Pract* 2005;**10**:1–3.
8 Streiner DL, Norman GR. *Health Measurement Scales: A Practical Guide to their Development and Use*, 4th edn. New York, NY: Oxford University Press 2008.
9 Cunnington JPW, Neville AJ, Norman GR. The risks of thoroughness: reliability and validity of global ratings and checklists in an OSCE. *Adv Health Sci Educ Theory Pract* 1996;**1**:227–33.
10 Norman GR, van der Vleuten CPM, De Graaff E. Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. *Med Educ* 1991;**25**:119–26.
11 Hodges B, McIlroy JH. Analytic global OSCE ratings are sensitive to level of training. *Med Educ* 2003;**37**:1012–6.
12 Govaerts MB, van der Vleuten CM, Schuwirth LT, Muijtjens AM. Broadening perspectives on clinical performance assessment: rethinking the nature of in-training assessment. *Adv Health Sci Educ Theory Pract* 2007;**12**:239–60.
13 Eva KW, Hodges BD. Scylla or Charybdis? Can we navigate between objectification and judgement in assessment? *Med Educ* 2012;**46**:914–9.
14 Schuwirth LWT, van der Vleuten CPM. A plea for new psychometric models in educational assessment. *Med Educ* 2006;**40**:296–300.
15 Lievens F. Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *J Appl Psychol* 2001;**86**:255–64.
16 Holmboe ES, Hawkins RE, Huot SJ. Effects of training in direct observation of medical residents' clinical competence: a randomised trial. *Ann Intern Med* 2004;**140**:874–81.
17 Kogan JR, Hess BJ, Conforti LN, Holmboe ES. What drives faculty ratings of residents' clinical skills? The impact of faculty's own clinical skills. *Acad Med* 2010;**85** (Suppl):25–8.
18 Brannick MT, Erol-Korkmaz HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. *Med Educ* 2011;**45**:1181–9.
19 Khan KZ, Gaunt K, Ramachandran S, Pushkar P. The objective structured clinical examination (OSCE): AMEE Guide No. 81. Part II: organisation and administration. *Med Teach* 2013;**35**:1447–63.
20 Hettinga AM, Denessen E, Postma CT. Checking the checklist: a content analysis of expert- and evidence-based case-specific checklist items. *Med Educ* 2010;**44**:874–83.
21 Cook DA, Brydges R, Zendejas B, Hamstra SJ, Hatala R. Technology-enhanced simulation to assess health professionals: a systematic review of validity evidence, research methods, and reporting quality. *Acad Med* 2013;**88**:872–83.
22 Brydges R, Hatala R, Zendejas B, Erwin PJ, Cook DA. Linking simulation-based educational assessments and patient-related outcomes: a systematic review and meta-analysis. *Acad Med* 2014. Epub ahead of print November 4, 2014. doi: 10.1097/ACM.0000000000000549.
23 Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann Intern Med* 2009;**151**:264–9.
24 Cook DA, Hatala R, Brydges R, Zendejas B, Szostek JH, Wang AT, Erwin PJ, Hamstra SJ. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. *JAMA* 2011;**306**:978–88.
25 Messick S. Validity. In: Linn RL, ed. *Educational Measurement*, 3rd edn. New York, NY: American Council on Education and Macmillan 1989;13–103.
26 Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med* 2006;**119**:166.e7–16.
27 Reed DA, Cook DA, Beckman TJ, Levine RB, Kern DE, Wright SM. Association between funding and quality of published medical education research. *JAMA* 2007;**298**:1002–9.
28 Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003;**327**:557–60.
29 Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C, Brown M. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg* 1997;**84**:273–8.
30 Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;**86**:420–8.
31 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;**33**:159–74.
32 Nunnally JC. *Psychometric Theory*, 2nd edn. New York, NY: McGraw-Hill 1978.
33 Cohen J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn. Hillsdale, NJ: Lawrence Erlbaum 1988.
34 Murray DJ, Boulet JR, Kras JF, McAllister JD, Cox TE. A simulation-based acute skills performance assessment for anaesthesia training. *Anesth Analg* 2005;**101**:1127–34.

35 White MA, DeHaan AP, Stephens DD, Maes AA, Maatman TJ. Validation of a high fidelity adult ureteroscopy and renoscopy simulator. *J Urol* 2010;**183**:673–7.
36 Finan E, Bismilla Z, Campbell C, Leblanc V, Jefferies A, Whyte HE. Improved procedural performance following a simulation training session may not be transferable to the clinical environment. *J Perinatol* 2012;**32**:539–44.
37 Gordon JA, Alexander EK, Lockley SW, Flynn-Evans E, Venkatan SK, Landrigan CP, Czeisler CA, Harvard Work Hours Health and Safety Group. Does simulator-based clinical performance correlate with actual hospital behaviour? The effect of extended work hours on patient care provided by medical interns. *Acad Med* 2010;**85**:1583–8.
38 Mazor KM, Ockene JK, Rogers HJ, Carlin MM, Quirk ME. The relationship between checklist scores on a communication OSCE and analogue patients' perceptions of communication. *Adv Health Sci Educ Theory Pract* 2005;**10**:37–51.
39 Sackett PR, Laczo RM, Arvey RD. The effects of range restriction on estimates of criterion interrater reliability: implications for validation research. *Pers Psychol* 2002;**55**:807–25.
40 Holmboe ES, Ward DS, Reznick RK, Katsufrakis PJ, Leslie KM, Patel VL, Ray DD, Nelson EA. Faculty development in assessment: the missing link in competency-based medical education. *Acad Med* 2011;**86**:460–7.
41 Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. *Med Teach* 2010;**32**:676–82.
42 Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. Effect of rater training on reliability and accuracy of mini-CEX scores: a randomised, controlled trial. *J Gen Intern Med* 2009;**24**:74–9.
43 Schuwirth LWT, van der Vleuten CPM. Programmatic assessment and Kane's validity perspective. *Med Educ* 2012;**46**:38–48.
44 Schuwirth LW, van der Vleuten CP. Programmatic assessment: from assessment of learning to assessment for learning. *Med Teach* 2011;**33**:478–85.
45 Boulet JR, Murray D. Review article: assessment in anaesthesiology education. *Can J Anaesth* 2012;**59**:182–92.
46 Jansen JJ, Berden HJ, van der Vleuten CP, Grol RP, Rethans J, Verhoeff CP. Evaluation of cardiopulmonary resuscitation skills of general practitioners using different scoring methods. *Resuscitation* 1997;**34**:35–41.
47 Reznick R, Regehr G, MacRae H, Martin J, McCulloch W. Testing technical skill via an innovative 'bench station' examination. *Am J Surg* 1997;**173**:226–30.
48 Friedlich M, MacRae H, Oandasan I, Tannenbaum D, Batty H, Reznick R, Regehr G. Structured assessment of minor surgical skills (SAMSS) for family medicine residents. *Acad Med* 2001;**76**:1241–6.
49 Morgan PJ, Cleave-Hogg D, Guest CB. A comparison of global ratings and checklist scores from an undergraduate assessment using an anaesthesia simulator. *Acad Med* 2001;**76**:1053–5.
50 Murray D, Boulet J, Ziv A, Woodhouse J, Kras J, McAllister J. An acute care skills evaluation for graduating medical students: a pilot study using clinical simulation. *Med Educ* 2002;**36**:833–41.
51 Adrales GL, Park AE, Chu UB, Witzke DB, Donnelly MB, Hoskins JD, Mastrangelo MJ Jr, Gandsas A. A valid method of laparoscopic simulation training and competence assessment. *J Surg Res* 2003;**114**:156–62.
52 Datta V, Bann S, Beard J, Mandalia M, Darzi A. Comparison of bench test evaluations of surgical skill with live operating performance assessments. *J Am Coll Surg* 2004;**199**:603–6.
53 Murray DJ, Boulet JR, Kras JF, Woodhouse JA, Cox T, McAllister JD. Acute care skills in anaesthesia practice: a simulation-based resident performance assessment. *Anesthesiology* 2004;**101**:1084–95.
54 Weller J, Robinson B, Larsen P, Caldwell C. Simulation-based training to improve acute care skills in medical undergraduates. *N Z Med J* 2004;**117**:U1119.
55 Bann S, Davis IM, Moorthy K, Munz Y, Hernandez J, Khan M, Datta V, Darzi A. The reliability of multiple objective measures of surgery and the role of human performance. *Am J Surg* 2005;**189**:747–52.
56 Moorthy K, Munz Y, Adams S, Pandey V, Darzi A. A human factors analysis of technical and team skills among surgical trainees during procedural simulations in a simulated operating theatre. *Ann Surg* 2005;**242**:631–9.
57 Berkenstadt H, Ziv A, Gafni N, Sidi A. The validation process of incorporating simulation-based accreditation into the anaesthesiology Israeli national board exams. *Isr Med Assoc J* 2006;**8**:728–33.
58 Broe D, Ridgway PF, Johnson S, Tierney S, Conlon KC. Construct validation of a novel hybrid surgical simulator. *Surg Endosc* 2006;**20**:900–4.
59 Matsumoto ED, Pace KT, D'A Honey RJ. Virtual reality ureteroscopy simulator as a valid tool for assessing endourological skills. *Int J Urol* 2006;**13**:896–901.
60 Banks EH, Chudnoff S, Karmin I, Wang C, Pardanani S. Does a surgical simulator improve resident operative performance of laparoscopic tubal ligation? *Am J Obstet Gynecol* 2007;**197**:541.e1–5.
61 Fialkow M, Mandel L, Van Blaricom A, Chinn M, Lentz G, Goff B. A curriculum for Burch colposuspension and diagnostic cystoscopy evaluated by an objective structured assessment of technical skills. *Am J Obstet Gynecol* 2007;**197**:544.e1–6.
62 Goff BA, Van Blaricom A, Mandel L, Chinn M, Nielsen P. Comparison of objective, structured assessment of technical skills with a virtual reality hysteroscopy trainer and standard latex hysteroscopy model. *J Reprod Med* 2007;**52**:407–12.
63 Khan MS, Bann SD, Darzi AW, Butler PE. Assessing surgical skill using bench station models. *Plast Reconstr Surg* 2007;**120**:793–800.

64 Zirkle M, Taplin MA, Anthony R, Dubrowski A. Objective assessment of temporal bone drilling skills. *Ann Otol Rhinol Laryngol* 2007;**116**:793–8.
65 Leung RM, Leung J, Vescan A, Dubrowski A, Witterick I. Construct validation of a low-fidelity endoscopic sinus surgery simulator. *Am J Rhinol* 2008;**22**:642–8.
66 Siddiqui NY, Stepp KJ, Lasch SJ, Mangel JM, Wu JM. Objective structured assessment of technical skills for repair of fourth-degree perineal lacerations. *Am J Obstet Gynecol* 2008;**199**:676.e1–6.
67 Chipman JG, Schmitz CC. Using objective structured assessment of technical skills to evaluate a basic skills simulation curriculum for first-year surgical residents. *J Am Coll Surg* 2009;**209**:364–70.e2.
68 Huang GC, Newman LR, Schwartzstein RM, Clardy PF, Feller-Kopman D, Irish JT, Smith CC. Procedural competence in internal medicine residents: validity of a central venous catheter insertion assessment instrument. *Acad Med* 2009;**84**:1127–34.
69 Insel A, Carofino B, Leger R, Arciero R, Mazzocca AD. The development of an objective model to assess arthroscopic performance. *J Bone Joint Surg Am* 2009;**91**:2287–95.
70 LeBlanc VR, Tabak D, Kneebone R, Nestel D, MacRae H, Moulton CA. Psychometric properties of an integrated assessment of technical and communication skills. *Am J Surg* 2009;**197**:96–101.
71 Faulkner H, Regehr G, Martin J, Reznick R. Validation of an objective structured assessment of technical skill for residents. *Acad Med* 1996;**71**:1363–5.
72 Siddighi S, Kleeman SD, Baggish MS, Rooney CM, Pauls RN, Karram MM. Effects of an educational workshop on performance of fourth-degree perineal laceration repair. *Obstet Gynecol* 2007;**109**:289–94.
73 Adler MD, Vozenilek JA, Trainor JL, Eppich WJ, Wang EE, Beaumont JL, Aitchison PR, Pribaz PJ, Erickson T, Edison M, McGaghie WC. Comparison of checklist and anchored global rating instruments for performance rating of simulated paediatric emergencies. *Simul Healthc* 2011;**6**:18–24.
74 Tuchschmid S, Bajka M, Harders M. Comparing automatic simulator assessment with expert assessment of virtual surgical procedures. In: Bello F, Cotin S, eds. *Lecture Notes in Computer Science*, Vol. 5958. Berlin Heidelberg: Springer-Verlag 2010;181–91.
75 Ault G, Reznick R, MacRae H, Leadbetter W, DaRosa D, Joehl R, Peters J, Regehr G. Exporting a technical skills evaluation technology to other sites. *Am J Surg* 2001;**182**:254–6.
76 Khan MS, Bann SD, Darzi A, Butler PE. Use of suturing as a measure of technical competence. *Ann Plast Surg* 2003;**50**:304–8.
77 Ponton-Carss A, Hutchison C, Violato C. Assessment of communication, professionalism, and surgical skills in an objective structured performance-related examination (OSPRE): a psychometric study. *Am J Surg* 2011;**202**:433–40.
78 Fleming J, Kapoor K, Sevdalis N, Harries M. Validation of an operating room immersive microlaryngoscopy simulator. *Laryngoscope* 2012;**122**:1099–103.
79 Hall AK, Pickett W, Dagnone JD. Development and evaluation of a simulation-based resuscitation scenario assessment tool for emergency medicine residents. *CJEM* 2012;**14**:139–46.
80 Jabbour N, Reihsen T, Payne NR, Finkelstein M, Sweet RM, Sidman JD. Validated assessment tools for paediatric airway endoscopy simulation. *Otolaryngol Head Neck Surg* 2012;**147**:1131–5.
81 Ma IW, Zalunardo N, Pachev G, Beran T, Brown M, Hatala R, McLaughlin K. Comparing the use of global rating scale with checklists for the assessment of central venous catheterisation skills using simulation. *Adv Health Sci Educ Theory Pract* 2012;**17**:457–70.
82 Nimmons GL, Chang KE, Funk GF, Shonka DC, Pagedar NA. Validation of a task-specific scoring system for a microvascular surgery simulation model. *Laryngoscope* 2012;**122**:2164–8.
83 VanHeest A, Kuzel B, Agel J, Putnam M, Kalliainen L, Fletcher J. Objective structured assessment of technical skill in upper extremity surgery. *J Hand Surg Am* 2012;**37**:332–7.
84 Cicero MX, Riera A, Northrup V, Auerbach M, Pearson K, Baum CR. Design, validity, and reliability of a paediatric resident JumpSTART disaster triage scoring instrument. *Acad Pediatr* 2013;**13**:48–54.

SUPPORTING INFORMATION

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

Figure S1. Meta-analysis of correlation coefficients between global rating scales and checklists.
Figure S2. Meta-analysis of inter-rater reliability of global rating scales and checklists.
Figure S3. Meta-analysis of inter-item reliability of global rating scales and checklists.
Figure S4. Meta-analysis of inter-station reliability of global rating scales and checklists.
Appendix S1. Inter-rater reliability for data abstraction by study investigators.
Appendix S2. Trial flow diagram.
Appendix S3. Methodological quality, as evaluated by the Medical Education Research Study Quality Instrument (MERSQI).
Appendix S4. Comparisons of validity evidence for relations to other variables between global rating scales (GRS) and checklists (CL).

Received 29 May 2014; editorial comments to author 1 August 2014; accepted for publication 9 September 2014