REVIEW

Technology-Enhanced Simulation for Health Professions Education

A Systematic Review and Meta-analysis

David A. Cook, MD, MHPE
Rose Hatala, MD, MSc
Ryan Brydges, PhD
Benjamin Zendejas, MD, MSc
Jason H. Szostek, MD
Amy T. Wang, MD
Patricia J. Erwin, MLS
Stanley J. Hamstra, PhD

Context Although technology-enhanced simulation has widespread appeal, its effectiveness remains uncertain. A comprehensive synthesis of evidence may inform the use of simulation in health professions education.

Objective To summarize the outcomes of technology-enhanced simulation training for health professions learners in comparison with no intervention.

Data Source Systematic search of MEDLINE, EMBASE, CINAHL, ERIC, PsychINFO, Scopus, key journals, and previous review bibliographies through May 2011.

Study Selection Original research in any language evaluating simulation compared with no intervention for training practicing and student physicians, nurses, dentists, and other health care professionals.

Data Extraction Reviewers working in duplicate evaluated quality and abstracted information on learners, instructional design (curricular integration, distributing training over multiple days, feedback, mastery learning, and repetitive practice), and outcomes. We coded skills (performance in a test setting) separately for time, process, and product measures, and similarly classified patient care behaviors.

Data Synthesis From a pool of 10 903 articles, we identified 609 eligible studies enrolling 35 226 trainees. Of these, 137 were randomized studies, 67 were nonrandomized studies with 2 or more groups, and 405 used a single-group pretest-posttest design. We pooled effect sizes using random effects. Heterogeneity was large (I^2>50%) in all main analyses. In comparison with no intervention, pooled effect sizes were 1.20 (95% CI, 1.04-1.35) for knowledge outcomes (n=118 studies), 1.14 (95% CI, 1.03-1.25) for time skills (n=210), 1.09 (95% CI, 1.03-1.16) for process skills (n=426), 1.18 (95% CI, 0.98-1.37) for product skills (n=54), 0.79 (95% CI, 0.47-1.10) for time behaviors (n=20), 0.81 (95% CI, 0.66-0.96) for other behaviors (n=50), and 0.50 (95% CI, 0.34-0.66) for direct effects on patients (n=32). Subgroup analyses revealed no consistent statistically significant interactions between simulation training and instructional design features or study quality.

Conclusion In comparison with no intervention, technology-enhanced simulation training in health professions education is consistently associated with large effects for outcomes of knowledge, skills, and behaviors and moderate effects for patient-related outcomes.

JAMA. 2011;306(9):978-988 www.jama.com

RESPONDING TO CHANGING practice environments requires new models for training health care professionals. Technology-enhanced simulation is one possible solution. We define *technology* broadly as materials and devices created or adapted to solve practical problems. *Simulation technologies* encompass diverse products including computer-based virtual reality simulators, high-fidelity and static mannequins, plastic models, live animals, inert animal products, and human cadavers.

Although technology-enhanced simulation has widespread appeal and many assert its educational utility,[1] such beliefs presently lack empirical support. Despite the large volume of research on simulation, its effectiveness remains uncertain in part because of the difficulty in interpreting research results one study at a time. Several systematic reviews[2-5] and at least 2 meta-analyses[6,7] have attempted to provide such syntheses, but each had limitations, including narrow inclusion criteria, incomplete

Author Affiliations: Office of Education Research, Mayo Medical School (Dr Cook), and Division of General Internal Medicine (Drs Cook, Szostek, and Wang), Department of Surgery (Dr Zendejas), and Mayo Libraries (Ms Erwin), Mayo Clinic College of Medicine, Rochester, Minnesota; Department of Medicine, University of British Columbia, Vancouver, Canada (Dr Hatala); Department of Medicine, University of Toronto, Toronto, Ontario, Canada (Dr Brydges); and Academy for Innovation in Medical Education, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada (Dr Hamstra).
Corresponding Author: David A. Cook, MD, MHPE, Division of General Internal Medicine, Mayo Clinic College of Medicine, Baldwin 4-A, 200 First St SW, Rochester, MN 55905 (cook.david33@mayo.edu).

accounting of existing studies, limited assessment of study quality, or no quantitative pooling to derive best estimates of the effect of these interventions on educational and patient outcomes. We therefore sought to identify and quantitatively summarize all studies of technology-enhanced simulation involving health professions learners.

METHODS

This review was planned, conducted, and reported in adherence to PRISMA standards of quality for reporting meta-analyses.[8]

Study Questions

We sought to answer 2 questions: (1) To what extent are simulation technologies for training health care professionals associated with improved outcomes in comparison with no intervention? and (2) How do outcomes vary for different simulation instructional designs? Based on the strength of the theoretical foundations and currency in the field, we selected 5 instructional design features[2,9] (curricular integration, distributed practice, feedback, mastery learning, and range of difficulty) for subgroup analyses (see eBox for definitions; available at http://www.jama.com).

Study Eligibility

Broad inclusion criteria were used to present a comprehensive overview of technology-enhanced simulation in health professions education. Studies published in any language were included if they investigated use of technology-enhanced simulation to teach health professions learners at any stage in training or practice, in comparison with no intervention (ie, a control group or preintervention assessment), using outcomes[10] of learning (knowledge or skills in a test setting), behaviors (in practice), or effects on patients (eBox). We included single-group pretest-posttest, 2-group nonrandomized, and randomized studies; parallel-group and crossover designs; and studies of "adjuvant" instruction in which simulation was added to other instruction common to all learners.

Studies that evaluated computer-based virtual patients requiring only standard computer equipment were excluded because these have been the subject of a recent systematic review.[11] Studies that involved human patient actors (standardized patients) or simulation for noneducation purposes such as procedural planning, disease modeling, or evaluating the outcomes of clinical interventions were also excluded.

Study Identification

An experienced research librarian (P.J.E.) designed a strategy (eAppendix) to search MEDLINE, EMBASE, CINAHL, PsychINFO, ERIC, Web of Science, and Scopus using search terms for the intervention (eg, *simulator*, *simulation*, *manikin*, *cadaver*, *MIST*, *Harvey*), topic (eg, *surgery*, *endoscopy*, *anesthesia*, *trauma*, *colonoscopy*), and learners (eg, *education medical*, *education nursing*, *education professional*, *students health occupations*, *internship and residency*). No beginning date cutoff was used, and the last date of search was May 11, 2011. This search was supplemented by adding the entire reference lists from several published reviews of health professions simulation and all articles published in 2 journals devoted to health professions simulation (*Simulation in Healthcare* and *Clinical Simulation in Nursing*) since their inception. Additional studies were sought from authors' files. We searched for additional studies in the reference lists of all included articles published before 1990 and a random sample of 160 included articles published in or after 1990.

Study Selection

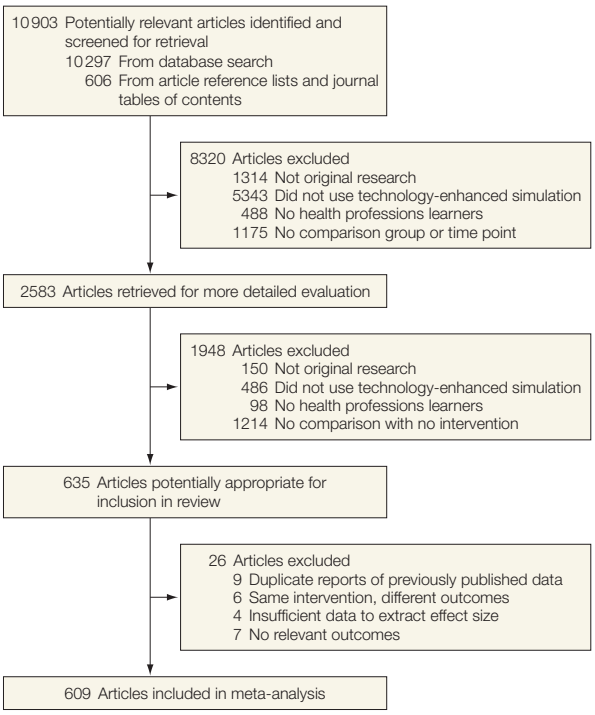
The authors worked independently and in duplicate to screen all titles and abstracts for inclusion. In the event of disagreement or insufficient information in the abstract, the full text of potential articles was reviewed independently and in duplicate. Conflicts were resolved by consensus. Chance-adjusted interrater agreement for study inclusion, determined using the intraclass correlation coefficient[12] (ICC), was 0.69.

Data Extraction

A data abstraction form was developed through iterative testing and revision. Data were extracted independently and in duplicate for all variables when reviewer judgment was required, using the ICC to determine interrater agreement and resolving conflicts by consensus.

We abstracted information on the training level of learners, clinical topic, training location (simulation center or clinical environment), study design, method of group assignment, outcomes, and methodological quality. Methodological quality was graded using the Medical Education Research Study Quality Instrument (MERSQI)[13] and an adaptation of the Newcastle-Ottawa Scale (NOS) for cohort studies[14,15] that evaluates representativeness of the intervention group (ICC, 0.86), selection of the comparison group (ICC, 0.29, with 86% raw agreement), comparability of cohorts (statistical adjustment for baseline characteristics in nonrandomized studies [ICC, 0.86] or randomization [ICC, 0.89] and allocation concealment for randomized studies [ICC, 0.65]), blinding of outcome assessment (ICC, 0.58), and completeness of follow-up (ICC, 0.36, with 78% raw agreement). We further coded the presence of simulation features identified in a review of simulation[2]: feedback (ICC, 0.47), repetitive practice (ICC, 0.42), curriculum integration (ICC, 0.53), range of task difficulty (ICC, 0.37), multiple learning strategies (ICC, 0.45), clinical variation (ICC, 0.49), and individualized learning (ICC, 0.31). We also coded the presence of mastery learning (ICC, 0.65), distributed practice (whether learners trained on 1 or >1 day; ICC, 0.73), and cognitive interactivity (ICC, 0.28), the duration of training (ICC, 0.68), and the number of task repetitions (ICC, 0.69). We initially planned to abstract informa-

Figure 1. Study Flow



tion on simulation fidelity but were unable to operationalize this construct with high reliability.

Outcomes were distinguished using the Kirkpatrick classification[10] and information was abstracted separately for learning (knowledge and skills, with skill measures further classified as time, process, and product), behaviors with patients (time and process measures), and results (patient effects). When authors reported multiple measures of a single outcome (eg, multiple measures of efficiency), we selected in decreasing order of priority (1) the author-defined primary outcome; (2) a global or summary measure of effect; (3) the most clinically relevant measure; or (4) the mean of the measures reported. We also prioritized skill outcomes assessed in a different setting (eg, different simulator or clinical setting) over those assessed via the simulator used for training. When abstracting data from learning curves, the first time point was used as the pretest and the last reported time point as the postintervention assessment.

Data Synthesis

Each mean and standard deviation or odds ratio was converted to a standardized mean difference (the Hedges *g* effect size).[16-18] When this information was unavailable, the effect size was estimated using statistical test results (eg, *P* values).[16] For 2-group pretest-posttest studies, we used posttest means adjusted for pretest or adjusted statistical test results; if these were not available, we standardized the difference in change scores using the pretest variance.[17] For crossover studies, we used means or exact statistical test results adjusted for repeated measures; if these were not available, we used means pooled across each intervention.[19,20] For studies reporting neither *P* values nor any measure of variance, we used the average standard deviation from all other studies reporting that outcome. If articles contained insufficient information to calculate an effect size, we requested this information from authors via e-mail.

The I^2 statistic[21] was used to quantify inconsistency (heterogeneity) across studies, with values greater than 50% indicating high inconsistency. Because there was high inconsistency in most analyses, random-effects models were used to pool weighted effect sizes. Planned subgroup analyses were conducted based on study design (randomized vs nonrandomized), total quality score, and selected instructional design features (presence of curricular integration, distributed practice over >1 day, feedback, mastery learning, and range of task difficulty) using the *z* test[22] to evaluate the statistical significance of interactions. As thresholds for high or low quality scores, we used an NOS score of 4 (as described previously[15]) and the median of the MERSQI scores (12). Sensitivity analyses were performed excluding studies that used *P* value upper limits or imputed standard deviations to estimate the effect size. Although funnel plots can be misleading in the presence of inconsistency,[23] we used these along with the Egger asymmetry test[24] to explore possible publication bias. In cases of asymmetry, trim and fill was used to estimate revised pooled effect size estimates, although this method also has limitations when inconsistency is present.[25] SAS software, version 9.1 (SAS Institute Inc, Cary, North Carolina) was used for analyses. Statistical significance was defined by a 2-sided $\alpha = .05$, and interpretations of clinical significance emphasized confidence intervals in relation to Cohen effect size classifications (>0.8 = large; 0.5-0.8 = moderate).[26]

RESULTS

Trial Flow

We identified 10 903 potentially relevant articles, 10 297 using the search strategy and 606 from the review of reference lists and journal indexes. From these, we identified 635 studies comparing simulation training with no intervention (FIGURE 1), of which 628 reported an eligible outcome. We identified 9 reports of previously reported data and 6 articles reporting different outcomes for previously de-

scribed interventions; for these, we selected the most detailed report for full review. Nine articles contained insufficient data to calculate an effect size. We received additional information from 5 authors; the remaining 4 were excluded. Ultimately, we analyzed 609 studies enrolling 35 226 trainees. TABLE 1 summarizes key study features and eTable 1 lists all references with additional information.

Study Characteristics

Since the earliest study we identified, published in 1969,[27] investigators have evaluated the use of technology-enhanced simulations in teaching laparoscopic surgery, gastrointestinal endoscopy, suturing skills, emergency resuscitation, team leadership, examination of the heart, breast, and pelvis, and many other topics. Nearly half the articles (n=282) were published in or after 2008, and 24 were published in a language other than English. Learners in these studies encompass a broad range of health care professionals, including physicians, nurses, emergency medicine technicians, military medics, dentists, chiropractors, veterinarians, and other allied health staff, and range from novice to expert.

Of the 609 studies, 274 spread training across more than 1 day, 108 provided high feedback, and 59 used a mastery learning model. TABLE 2 lists other simulation key features. Of 910 outcomes reported in these studies, 785 (86%) were objectively determined (eg, by faculty ratings or computer scoring). The majority (n=690) assessed skills in a training setting, including time to complete the task, process measures (eg, global ratings of performance, economy of movements in surgery, or minor errors), and task products (eg, quality of a dental preparation, procedural success, failure to detect key abnormalities, or major procedural complication). Skills were usually assessed with the simulator that was used for training. However, 36 of 210 time outcome measures (17%), 58 of 426 process measures (14%), and 15 of 54 product measures (28%) assessed skills using another simulation mode (ie, skill transfer). Knowledge outcomes (usually multiple-choice test scores) were reported in 118 studies; time and process behaviors with real patients in 20 and 50 studies, respec-

Table 1. Characteristics of Included Studies[a]

Study Characteristics	No. of Studies	No. of Participants[b]
All studies	609	35 226
Study design		
Posttest-only, 2 groups	110	8362
Pretest-posttest, 2 groups	94	3784
Pretest-posttest, 1 group	405	23 080
Randomized group allocation	137	5093
Location		
Simulation center	564	30 551
Clinical environment	34	3771
Both simulation center and clinical environment	11	904
Participants[c]		
Medical students	156	9530
Physicians in postgraduate training	324	8712
Physicians in practice	135	5690
Nurses and nursing students	79	4146
Emergency medical technicians and students	20	1198
Dentists and dental students	12	624
Veterinarians and veterinary students	6	145
Chiropractors and chiropractic students	1	60
Other/ambiguous/mixed	79	5121
Clinical topics[c,d]		
Minimally invasive surgery	158	4741
Resuscitation/trauma training	87	7330
Endoscopy and ureteroscopy	72	1861
Other surgery	66	4819
Physical examination	37	4653
Intubation	35	2197
Communication and team skills	33	2451
Vascular access	31	2264
Obstetrics	25	1774
Anesthesia	23	2134
Endovascular procedures	10	201
Dentistry	9	395
Outcomes[c]		
Knowledge	118	8595
Time skills	210	5651
Process skills	426	20 926
Product skills	54	2158
Time behaviors	20	384
Process behaviors	50	1975
Patient effects	32	1648
Quality		
Medical Education Research Study Quality Instrument ≥12 points	295	14 849
Newcastle-Ottawa Scale ≥4 points	111	4513

[a] See eTable 1 for details on individual studies.
[b] Numbers reflect the number of participants enrolled, except for outcomes, which reflect the number of participants who provided observations for analysis.
[c] The number of studies and learners in some subgroups may sum to more than the number for all studies and percentages may total more than 100% because several studies included more than 1 learner group, addressed more than 1 clinical topic, or reported multiple outcomes.
[d] Selected listing of the topics addressed most often (numerous other topics were addressed with lower frequency).

Table 2. Prevalence of Simulation Key Features

Features	Studies, No. (%) (N = 609)
Blended learning (nonsimulation activities) (high)	181 (29.7)
Clinical variation (present)	193 (31.7)
Cognitive interactivity (high)	359 (58.9)
Curriculum integration (present)	102 (16.7)
Distributed practice (>1 d)	274 (45)
Feedback (high)	108 (17.7)
Individualized learning (high)	26 (4.3)
Mastery learning (present)	59 (9.7)
Multiple learning strategies (high)	56 (9.2)
Range of task difficulty (present)	93 (15.3)
Repetitive practice (present)	484 (79.5)

tively; and direct patient effects in 32 studies. Behavior process measures included instructor ratings of competence, completion of key procedural elements, and procedural errors, while patient effects included procedural success, patient discomfort, complication rate, and patient survival; eTable 2 lists all behavior and patient outcomes.

Study Quality

TABLE 3 summarizes the methodological quality of included studies. The number of participants providing outcomes ranged from 2 to 1333, with a median of 24 (interquartile range, 15-47). One hundred thirty-seven of 204 2-group studies (67%) were randomized. Three 2-group studies (1.5%) determined groups by self-selection or completion/noncompletion of training. Although such groupings are susceptible to bias, sensitivity analyses excluding these studies showed similar results. Twenty-four of 118 studies (20%) assessing knowledge, 73 of 210 (35%) assessing time skills, 133 of 426 (31%) assessing process skills, 18 of 54 (33%) assessing product skills, 10 of 20 (50%) assessing time behaviors, 12 of 50 (24%) assessing process behaviors, and 12 of 32 (38%) assessing patient effects lost more than 25% of participants from time of enrollment or failed to report follow-up. Again, sensitivity analyses excluding these studies showed no substantial difference. Assessors were blinded to the study intervention for 503 outcome measures (55%), but only for knowledge outcomes did blinding show a significant association with effect size. Mean quality scores were relatively low, averaging 2.1 (SD, 1.5) for the NOS (maximum, 6 points) and 11.6 (SD, 1.9) for the MERSQI (maximum, 18 points).

Meta-analysis

FIGURE 2, FIGURE 3, FIGURE 4, and FIGURE 5 (also eFigures 1-7) summarize the meta-analysis results. In general, simulation training was associated with moderate to large, statistically significant positive results but with high inconsistency. Subgroup analyses demonstrated no consistent interactions. Sensitivity analyses did not alter study conclusions, and where funnel plots were asymmetric, trim-and-fill analyses yielded results similar to the original.

Knowledge. One hundred eighteen studies (with 8595 participants providing data) reported comparison with a preintervention assessment or a no-intervention control group using knowledge as the outcome (Figure 2a and eFigure 1). The pooled effect size for these interventions was 1.20 (95% CI, 1.04-1.35; $P<.001$), consistent with large[26] gains. However, there was high inconsistency among studies, with individual effect sizes ranging from –0.42 to 9.45 and $I^2=96\%$. Three studies (studies 70, 365, and 416 in eTable 1) reported a negative effect size (ie, outcomes were worse for simulation), although in the study with the largest negative effect size (study 70), the skills outcomes showed substantial benefit. The funnel plot was asymmetric and the Egger asymmetry test result was significant. Assuming this asymmetry reflects publication bias, trim-and-fill analyses provided a smaller but still large pooled effect size of 0.86 (95% CI, 0.68-1.04). Among the 16 studies with randomized group assignment, the pooled effect size was 0.63.

Tests for interactions in subgroups indicated that interventions distributed over more than 1 day (vs a single day) and those that were optional (vs integrated or required activities) were associated with significantly larger effect sizes. Studies scoring low on the modified NOS were associated with larger effect sizes than high-quality studies, and blinded outcomes were associated with higher effect sizes than unblinded assessments.

Time Skills. Two hundred ten studies (5651 participants) reported the time required to complete the task in a simulation setting (Figure 2b and eFigure 2). The pooled effect size of 1.14 (95% CI, 1.03-1.25; $P<.001$) reflects a large favorable association. There was high inconsistency across studies ($I^2=84\%$), and effect sizes ranged from –1.55 to 21.0. In each of the 10 studies (studies 171, 176, 215, 269, 270, 292, 354, 367, 407, and 525 in eTable 1) with negative effect sizes for time skills, other skill outcomes showed substantial improvements (ie, learners took longer but performed better). Funnel plot analysis suggested possible publication bias, but trim-and-fill analysis decreased the effect size only slightly to 1.10 (95% CI, 0.98-1.21). The 47 randomized trials had a pooled effect size of 0.75.

Interventions providing low (vs high) feedback and those with curricular integration were associated with larger improvements in time outcomes. Lower study quality, as measured by both the NOS and MERSQI, was associated with significantly larger effect sizes.

Process Skills. Four hundred twenty-six studies (20 926 participants) reported skill measures of process (eg, global ratings or efficiency) (Figure 3a and eFigure 3). The pooled effect size of 1.09 (95% CI, 1.03-1.16; $P<.001$) reflects a large favorable association but large inconsistency ($I^2=89\%$). Effect sizes ranged from –0.50 to 8.55. Possible explanations for the 7 studies with negative effect sizes (studies 95, 126, 185, 193, 319, 496, and 566 in eTable 1) include misalignment be-

tween the intervention and the assessment (2 cases [studies 95 and 185] of training for one procedure [eg, cholecystectomy] and assessing performance on a different procedure [eg, appendectomy]) and delayed testing with continued clinical training during the interim (a 3-month [study 566] or 2-year [study 126] delay). In 2 other cases (studies 319 and 496), all other outcomes assessed showed improvements. Funnel plot analysis suggested possible publication bias; trim-and-fill analysis provided a smaller but still large pooled effect size of 0.94 (95% CI, 0.87-1.01). The 87 randomized trials had a pooled effect size of 0.98.

Interventions using a mastery learning model, in which learners must achieve a rigorously defined benchmark before proceeding, were associated with significantly higher outcomes than nonmastery interventions.

Product Skills. Fifty-four studies (2158 participants) evaluated the products of the learners' performance, such as procedural success or the quality of a finished product (Figure 3b and eFigure 4). There was a large pooled effect size of 1.18 (95% CI, 0.98-1.37; $P<.001$) and high inconsistency ($I^2=87\%$). Effect sizes ranged from –0.09 to 6.24. For the single study (study 58 in eTable 1) showing a negative product skills effect size, other skill outcomes showed a positive association. Trim-and-fill analysis in response to an asymmetric funnel plot yielded a pooled effect size of 1.0 (95% CI, 0.77-1.23). The 15 randomized trials had a pooled effect size of 0.76.

Interventions using a mastery learning model were associated with lower learning outcomes, and lower study quality was once again associated with significantly higher learning outcomes.

Behavior. Twenty studies (384 participants) used a measure of time to evaluate behaviors while caring for patients (Figure 4a and eFigure 5). The association approached a large effect size (0.79; 95% CI, 0.47-1.10; $P<.001$), with $I^2=66\%$. Effect sizes ranged from –0.24 to 5.6, and for each of 3 studies with negative effect sizes (studies 151, 504, and 530 in eTable 1), the other behavior outcomes showed large positive associations. The funnel plot was reasonably symmetric. In contrast with other outcomes, higher study quality was associated with larger effect sizes. The 13 randomized trials had a pooled effect size of 1.01.

Fifty studies (1975 participants) reported other learner behaviors while caring for patients (Figure 4b and eFigure 6), with a large pooled effect size of 0.81 (95% CI, 0.66-0.96; $P<.001$) and $I^2=70\%$. Effect sizes ranged from –0.46 to 2.25. Three studies (studies 312, 547, and 583) found negative results. In one (study 312), the authors questioned the accuracy with which their simulator mimicked reality. Another (study 547) found slightly smaller behavior and patient effect outcomes when adding a 30-minute refresher course to routine intubation training. The 30 randomized trials had a pooled effect size of 0.85. There were no statistically significant as-

Table 3. Quality of Included Studies

Quality Measure (Points)	Studies, No. (%) (N = 609)
MERSQI score (maximum, 18)[a]	
Study design (maximum, 3)	
Pretest-posttest, 1 group (1.5)	405 (66.5)
Observational, 2 groups (2)	67 (11)
Randomized, 2 groups (3)	137 (22.5)
No. of institutions sampled (maximum, 1.5)	
1 (0.5)	517 (84.9)
2 (1)	24 (3.9)
>2 (1.5)	68 (11.2)
Follow-up, % (maximum, 1.5)	
<50 or not reported (0.5)	166 (27.3)
50-74 (1)	31 (5.1)
≥75 (1.5)	412 (67.6)
Outcome assessment (maximum, 3)	
Subjective (1)	90 (14.8)
Objective (3)	519 (85.2)
Validity evidence (maximum, 3)	
Content (1)	195 (32.0)
Internal structure (1)	130 (21.4)
Relation to other variables (1)	119 (19.5)
Data analysis	
Appropriate (maximum, 1)	527 (86.5)
Sophistication (maximum, 2)	
Descriptive (1)	49 (8.1)
Beyond descriptive analysis (2)	560 (91.9)
Highest outcome type (maximum, 3)	
Knowledge/skills (1.5)	544 (89.3)
Behaviors (2)	33 (5.4)
Patient/health care outcomes (3)	32 (5.3)
NOS (modified) score (maximum, 6)[b]	
Representativeness of sample (1)	134 (22.0)
Comparison group from same community (1)	190 (31.2)
Comparability of comparison cohort	
Criterion A (1)[c]	142 (23.3)
Criterion B (1)[c]	82 (13.5)
Blinded outcome assessment (1)	300 (49.3)
High rate of follow-up (1)	426 (70.0)

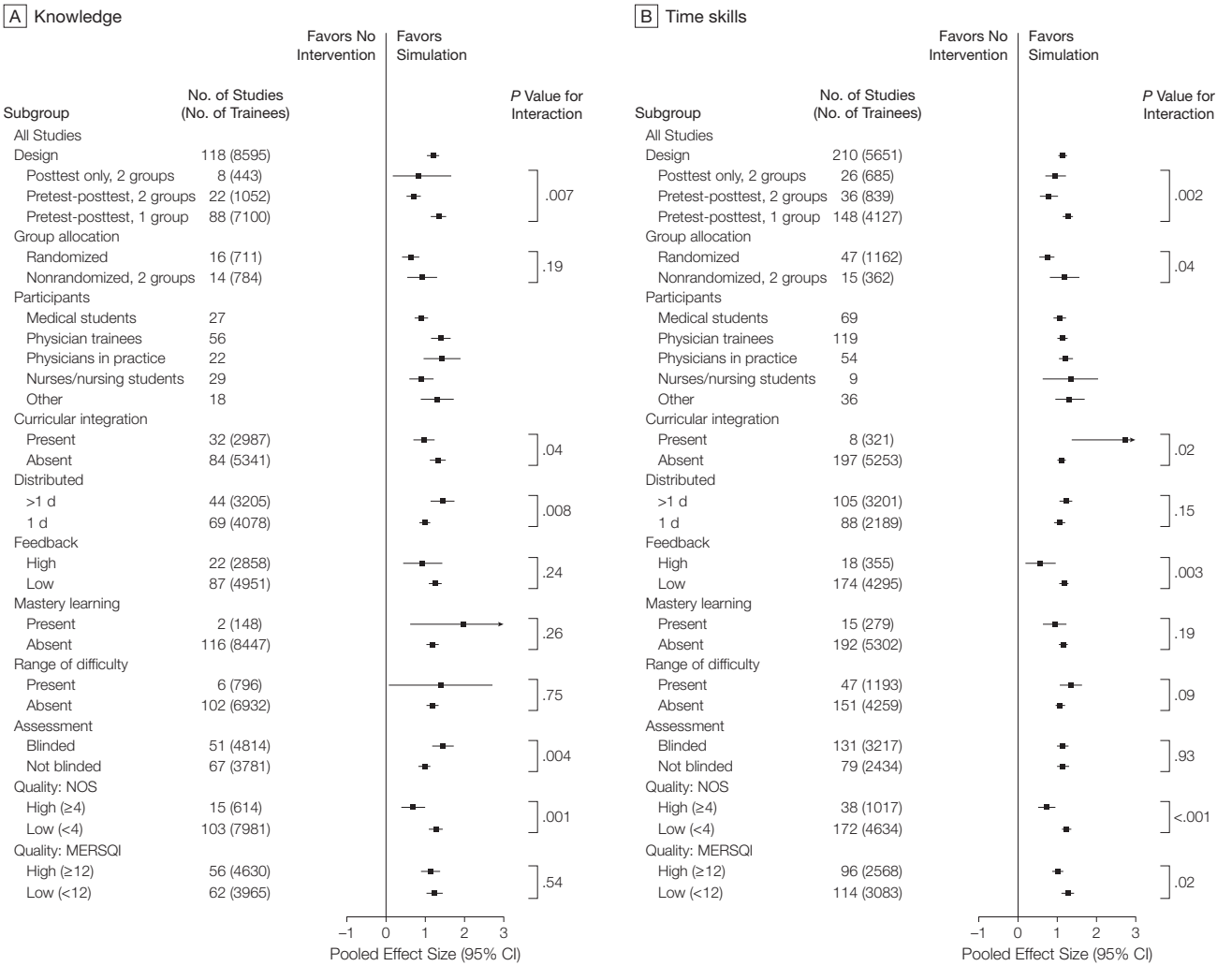
[a] The mean score on the Medical Education Research Study Quality Instrument (MERSQI) was 11.6 (SD, 1.9) and median score was 12 (range, 6.0-7.0).
[b] The mean score on the Newcastle-Ottawa Scale (NOS) was 2.1 (SD, 1.5) and median score was 2 (range, 0-6).
[c] Comparability of cohorts criterion A was fulfilled if the study (1) was randomized or (2) controlled for a baseline learning outcome. Criterion B was fulfilled if (1) a randomized study concealed allocation or (2) an observational study controlled for another baseline learner characteristic.

sociations in any of the planned subgroup analyses. The funnel plot was symmetric.

Effects on Patient Care. Thirty-two studies (1648 participants) reported effects on patient care (see aforementioned examples and Figure 5, eFigure 7, and eTable 2). For these outcomes, simulation training was associated with a moderate pooled effect size of 0.50 (95% CI, 0.34-0.66; $P<.001$). Inconsistency was high ($I^2=67\%$) and effect sizes ranged from –0.28 to 1.68, with 2 studies reporting a negative effect size. One of these studies (study 547) is described above; the other (study 387) found slightly worse patient outcomes but a substantial improvement in behaviors. The 14 randomized trials had a slightly smaller effect size of 0.37. The funnel plot was symmetric. Distributed training was associated with larger effect size.

Sensitivity Analyses. We used *P* value upper limits (eg, $P<.01$) to estimate 150 of 910 effect sizes (16%) and we imputed standard deviations to estimate 83 effect sizes (9%). Sensitivity analyses excluding these 233 effect sizes yielded pooled estimates similar to those of the full sample (approximately 5%-7% higher than those reported above for knowledge, process skills, and product skills outcomes; no change for patient effects; and 2%-7% lower for time skills, time behaviors, and other behaviors outcomes). Because most studies involved few participants, the possible dominance of very large studies was assessed by excluding the 22 studies with more than 200 participants; this analysis showed nearly identical results.

Figure 2. Random-Effects Meta-analysis of Simulation Training: Knowledge and Time Skills



Simulation compared with no intervention; positive numbers favor the simulation intervention. *P* values reflect statistical tests exploring the differential effect of simulation training (ie, interaction) for study subgroups. Participant groups are not mutually exclusive; thus, no statistical comparison is made and the number of trainees is not reported. Some features could not be discerned for all studies; hence, some subgroups do not sum to the total number of studies. NOS indicates Newcastle-Ottawa Scale; MERSQI, Medical Education Research Study Quality Instrument. See also eFigure 1 and eFigure 2.

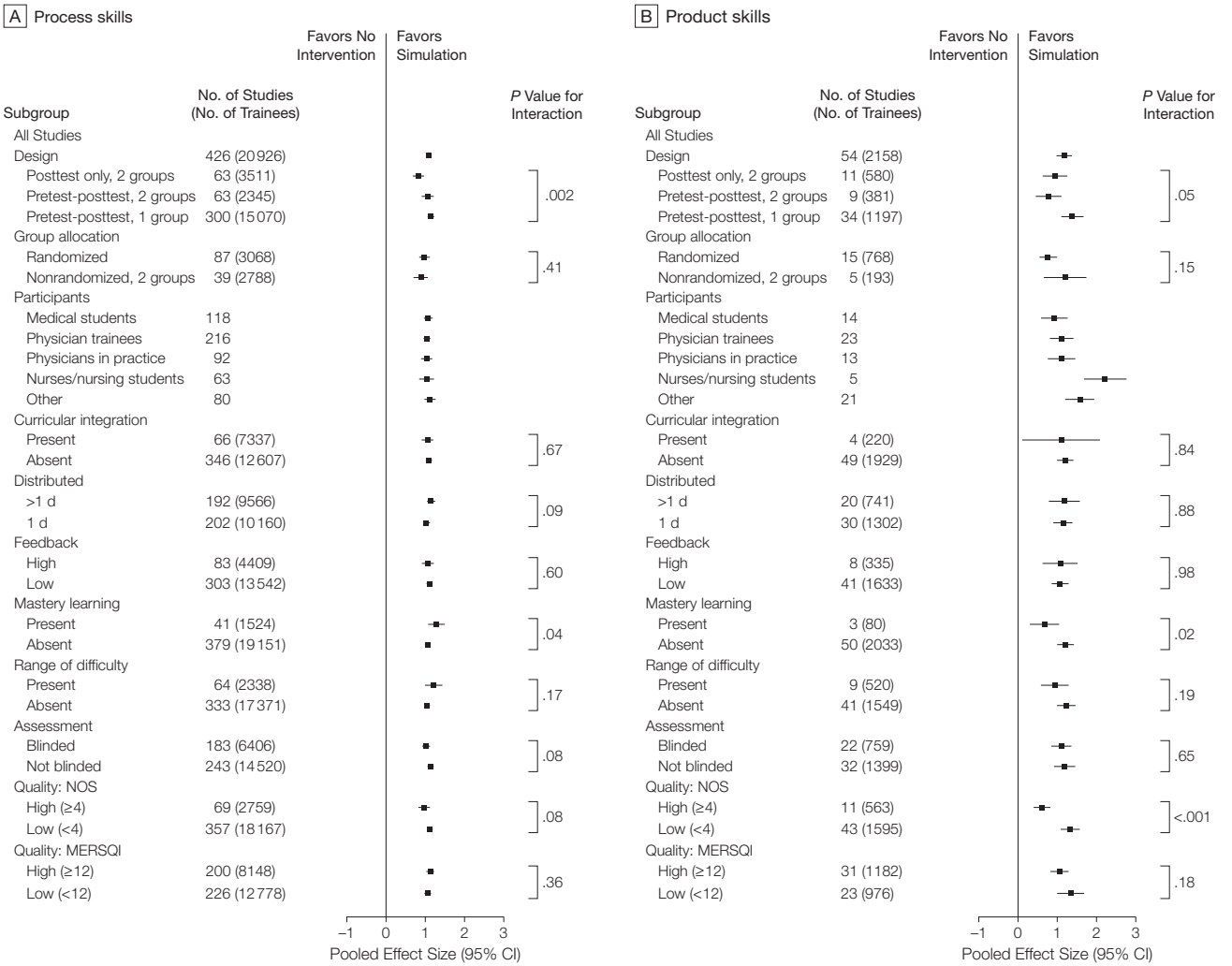
COMMENT

Technology-enhanced simulations, in comparison with no intervention or when added to traditional practice, were with only rare exceptions associated with better learning outcomes. Pooled effect sizes were large[26] for knowledge, skills, and behaviors, and confidence intervals excluded small associations. Effect sizes for patient-related outcomes were smaller but still moderate.

Yet in nearly all cases, the magnitude of association varied substantially for individual studies (high inconsistency), and subgroup analyses exploring simulation design differences largely failed to explain this variation. Although distributing learning activities over more than 1 day was consistently associated with larger effect sizes and mastery learning was associated with larger effect sizes for most outcomes, these differences were rarely statistically significant and inconsistency usually remained high.

In contrast, with rare exceptions there was lower inconsistency for 2-group posttest-only studies, randomized trials, and studies with high modified NOS scores. Also, 2-group studies and studies with high quality scores had a consistent (and often statistically significant) association with smaller effect sizes. These findings could be due to chance, other between-study differences such as variation in simulation design, or the sensitivity of the outcome measure. Still, it makes sense that studies with a comparison group, which helps control for maturation and learning outside of the in-

Figure 3. Random-Effects Meta-analysis of Simulation Training: Process and Product Skills



Simulation compared with no intervention; positive numbers favor the simulation intervention. *P* values reflect statistical tests exploring the differential effect of simulation training (ie, interaction) for study subgroups. Participant groups are not mutually exclusive; thus, no statistical comparison is made and the number of trainees is not reported. Some features could not be discerned for all studies; hence, some subgroups do not sum to the total number of studies. NOS indicates Newcastle-Ottawa Scale; MERSQI, Medical Education Research Study Quality Instrument. See also eFigure 3 and eFigure 4.

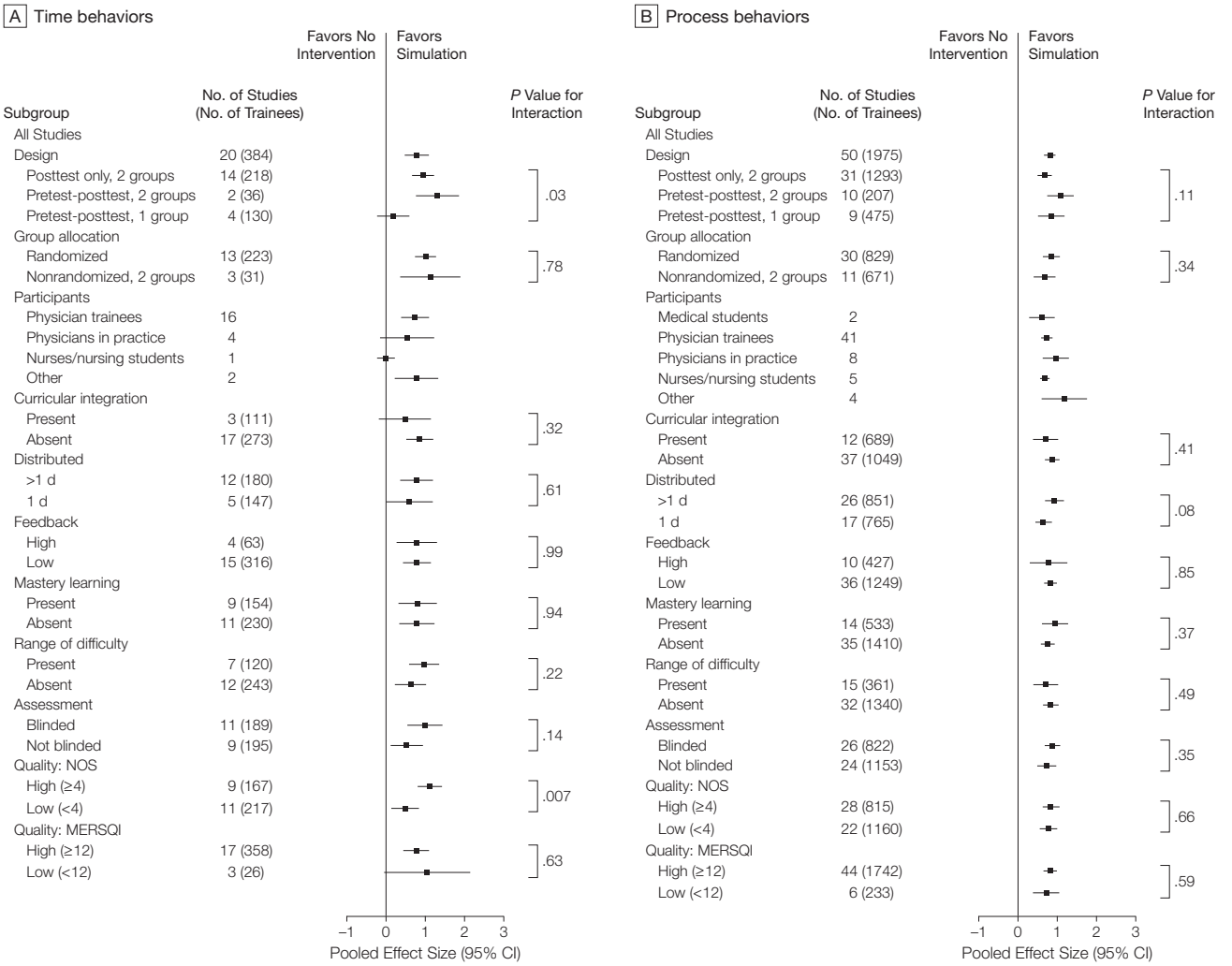
tervention, would show smaller effect sizes than single-group studies, and this has also been found in studies of Internet-based instruction.[28]

Limitations and Strengths

Because of this review's comprehensive scope, pooling outcomes across diverse simulation modes and educational contexts, we necessarily deferred exploring many topic- and design-specific interactions. The high inconsistency among studies was not surprising, reflecting variation not only in modes and clinical topics but also in learner groups, instructional designs, research methods, and outcome measures. Because the overwhelming majority of training interventions were associated with benefit, this heterogeneity could be interpreted as supporting the effectiveness of technology-enhanced simulation across a broad range of learners and topics.

Inferences are limited by the quality of available studies. Many reports failed to clearly describe the context, instructional design, or outcomes; poor reporting may have contributed to modest interrater agreement for some variables. Most studies had important methodological limitations. Two-thirds were single-group pretest-posttest comparisons and small samples were the norm. Most studies reported multiple measures of the same outcome, and although the results were usually congruent in direction, the magnitude varied slightly; however, we used a consistent approach in selecting measures to minimize bias. Because technology-enhanced simulations are

Figure 4. Random-Effects Meta-analysis of Simulation Training: Time and Process Behaviors



Simulation compared with no intervention; positive numbers favor the simulation intervention. *P* values reflect statistical tests exploring the differential effect of simulation training (ie, interaction) for study subgroups. Participant groups are not mutually exclusive; thus, no statistical comparison is made and the number of trainees is not reported. Some features could not be discerned for all studies; hence, some subgroups do not sum to the total number of studies. NOS indicates Newcastle-Ottawa Scale; MERSQI, Medical Education Research Study Quality Instrument. See also eFigure 5 and eFigure 6.

designed for health professions training, we did not include studies from nonhealth fields. We could not blind reviewers to article origins.

The subgroup analyses should be interpreted with caution because of the number of comparisons made, the absence of a priori hypotheses for many analyses, the limitations associated with between-study (rather than within-study) comparisons, and inconsistent findings across outcomes. For example, we found (contrary to expectation) that interventions with high feedback were often associated with smaller effect sizes. These and other counterintuitive results could be due to confounding (eg, simulation features unrelated to feedback that consistently affect the design or outcome), variation in outcome responsiveness, chance, or bias as well as true effect.

Strengths of the review include the exhaustive search; inclusion of multiple non-English articles; reproducible inclusion criteria encompassing a broad range of learners, outcomes, and study designs; duplicate, independent, and reproducible data abstraction; and rigorous coding of methodological quality. Funnel plots and trim-and-fill analyses suggested that publication bias is unlikely to affect our conclusions.

Comparison With Previous Reviews

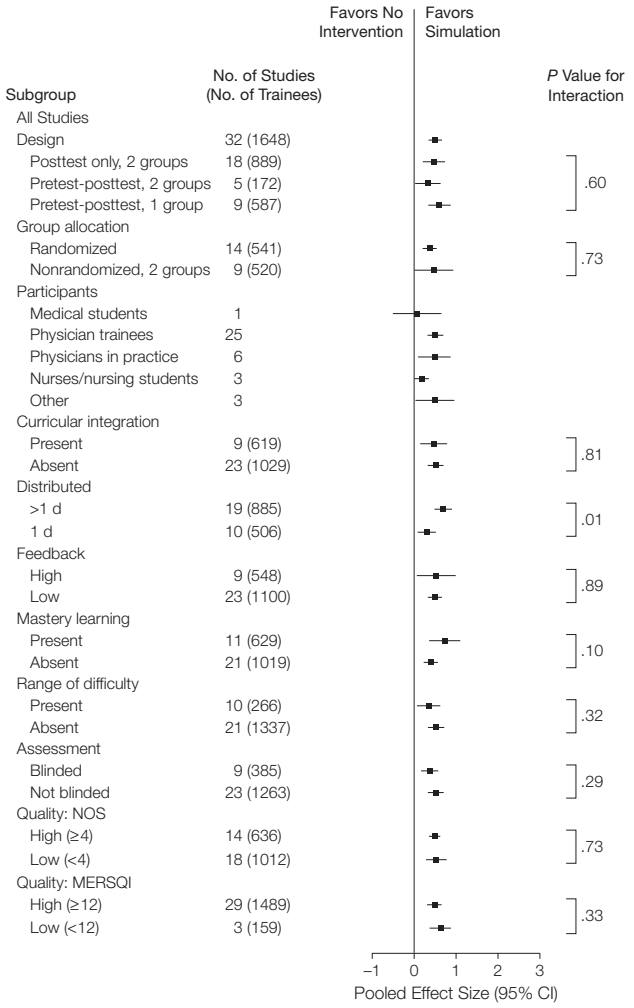
The 2005 review of high-fidelity simulation by Issenberg et al[2] identified 109 studies evaluating high-fidelity simulation in the health professions. Our meta-analysis contributes to the field by adding and synthesizing 500 additional studies. Previous meta-analyses of simulation training for health care professionals have found that laparoscopic surgery simulation (23 studies[6]) and training with deliberate practice (14 studies[7]) are associated with improved outcomes compared with no training. Systematic reviews of surgical simulation in general[4,5] have likewise concluded that simulation is beneficial. The finding of large associations when comparing technology-enhanced education with no intervention is consistent with recent meta-analyses of Internet-based instruction[15] and computer-based virtual patient simulations.[11]

Implications

Technology-enhanced simulation training is associated with improved outcomes in comparison with no intervention for health care professionals across a range of clinical topics and outcomes, including large effects on clinician behaviors and moderate effects on patient care. In light of such large associations, and with only 4% of the outcomes failing to show a benefit, we question the need for further studies comparing simulation with no intervention (ie, single-group pretest-posttest studies and comparisons with no-intervention controls).

The important questions for this field are those that clarify[29] when and how to use simulation most effectively and cost-efficiently.[30,31] Unfortunately, the

Figure 5. Random-Effects Meta-analysis of Simulation Training: Patient Effects



Simulation compared with no intervention; positive numbers favor the simulation intervention. *P* values reflect statistical tests exploring the differential effect of simulation training (ie, interaction) for study subgroups. Participant groups are not mutually exclusive; thus, no statistical comparison is made and the number of trainees is not reported. Some features could not be discerned for all studies; hence, some subgroups do not sum to the total number of studies. NOS indicates Newcastle-Ottawa Scale; MERSQI, Medical Education Research Study Quality Instrument. See also eFigure 7.

evidence synthesized herein largely fails to inform the design of future simulation activities. Subgroup analyses weakly suggested a benefit to extending training beyond 1 day and using a mastery model but otherwise did not identify consistent associations involving instructional designs. However, between-study (rather than within-study) comparisons are an inefficient research method.[32] Thus, theory-based comparisons between different technology-enhanced simulation designs (simulation vs simulation studies) that minimize bias, achieve appropriate power, and avoid confounding,[30] as well as rigorous qualitative studies, are necessary to clarify how and when to effectively use technology-enhanced simulations for training health care professionals.

Author Contributions: Dr Cook had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Study concept and design: Cook, Hatala, Hamstra.
Acquisition of data: Cook, Hatala, Brydges, Zendejas, Szostek, Wang, Erwin.
Analysis and interpretation of data: Cook.
Drafting of the manuscript: Cook.
Critical revision of the manuscript for important intellectual content: Cook, Hatala, Brydges, Zendejas, Szostek, Wang, Erwin, Hamstra.
Statistical analysis: Cook.
Administrative, technical, or material support: Cook, Zendejas, Wang, Erwin.
Study supervision: Cook.
Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.
Funding/Support: This work was supported by intramural funds, including an award from the Division of General Internal Medicine, Mayo Clinic.
Role of the Sponsor: The funding sources for this study played no role in the design and conduct of the study; in the collection, management, analysis, and interpretation of the data; or in the preparation of the manuscript. The funding sources did not review the manuscript.
Online-Only Material: The eBox, eAppendix, eTables 1 and 2, and eFigures 1 through 7 are available at http://www.jama.com.

REFERENCES

1. Gaba DM. The future vision of simulation in healthcare. *Simul Healthc*. 2007;2(2):126-135.
2. Issenberg SB, McGaghie WC, Petrusa ER, Lee Gordon D, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Med Teach*. 2005; 27(1):10-28.
3. McGaghie WC, Issenberg SB, Petrusa ER, Scalese RJ. Effect of practice on standardised learning outcomes in simulation-based medical education. *Med Educ*. 2006;40(8):792-797.
4. Sutherland LM, Middleton PF, Anthony A, et al. Surgical simulation: a systematic review. *Ann Surg*. 2006;243(3):291-300.
5. Sturm LP, Windsor JA, Cosman PH, Cregan P, Hewett PJ, Maddern GJ. A systematic review of skills transfer after surgical simulation training. *Ann Surg*. 2008;248(2):166-179.
6. Gurusamy K, Aggarwal R, Palanivelu L, Davidson BR. Systematic review of randomized controlled trials on the effectiveness of virtual reality training for laparoscopic surgery. *Br J Surg*. 2008;95(9):1088-1097.
7. McGaghie WC, Issenberg SB, Cohen ER, Barsuk JH, Wayne DB. Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? a meta-analytic comparative review of the evidence. *Acad Med*. 2011; 86(6):706-711.
8. 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(4):264-269, W64.
9. McGaghie WC, Issenberg SB, Petrusa ER, Scalese RJ. A critical review of simulation-based medical education research: 2003-2009. *Med Educ*. 2010; 44(1):50-63.
10. Kirkpatrick D. Revisiting Kirkpatrick's four-level model. *Train Dev*. 1996;50(1):54-59.
11. Cook DA, Erwin PJ, Triola MM. Computerized virtual patients in health professions education: a systematic review and meta-analysis. *Acad Med*. 2010; 85(10):1589-1602.
12. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979; 86(2):420-428.
13. 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(9):1002-1009.
14. Wells GA, Shea B, O'Connell D, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. http://www.ohri.ca/programs/clinical_epidemiology/oxford.htm. Accessed March 20, 2010.
15. Cook DA, Levinson AJ, Garside S, Dupras DM, Erwin PJ, Montori VM. Internet-based learning in the health professions: a meta-analysis. *JAMA*. 2008; 300(10):1181-1196.
16. Borenstein M. Effect sizes for continuous data. In: Cooper H, Hedges LV, Valentine JC, eds. *The Handbook of Research Synthesis.* 2nd ed. New York, NY: Russell Sage Foundation; 2009:221-235.
17. Morris SB, DeShon RP. Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychol Methods*. 2002; 7(1):105-125.
18. Hunter JE, Schmidt FL. *Methods of Meta-analysis: Correcting Error and Bias in Research Findings.* Thousand Oaks, CA: Sage; 2004.
19. Curtin F, Altman DG, Elbourne D. Meta-analysis combining parallel and cross-over clinical trials, I: continuous outcomes. *Stat Med*. 2002;21(15):2131-2144.
20. Higgins JPT, Green S. *Cochrane Handbook for Systematic Reviews of Interventions 5.1.0.* March 2011. http://www.cochrane.org/resources/handbook/index.htm. Accessed August 4, 2011.
21. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ*. 2003; 327(7414):557-560.
22. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Subgroup analyses. In: *Introduction to Meta-analysis.* Chichester, England: Wiley; 2009:chap 19.
23. Lau J, Ioannidis JPA, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. *BMJ*. 2006; 333(7568):597-600.
24. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ*. 1997;315(7109):629-634.
25. Terrin N, Schmid CH, Lau J, Olkin I. Adjusting for publication bias in the presence of heterogeneity. *Stat Med*. 2003;22(13):2113-2126.
26. Cohen J. *Statistical Power Analysis for the Behavioral Sciences.* 2nd ed. Hillsdale, NJ: Lawrence Erlbaum; 1988.
27. Abrahamson S, Denson JS, Wolf RM. Effectiveness of a simulator in training anesthesiology residents. *J Med Educ*. 1969;44(6):515-519.
28. Cook DA, Levinson AJ, Garside S. Method and reporting quality in health professions education research: a systematic review. *Med Educ*. 2011;45 (3):227-238.
29. Cook DA, Bordage G, Schmidt HG. Description, justification and clarification: a framework for classifying the purposes of research in medical education. *Med Educ*. 2008;42(2):128-133.
30. Cook DA. One drop at a time: research to advance the science of simulation. *Simul Healthc*. 2010; 5(1):1-4.
31. Weinger MB. The pharmacology of simulation: a conceptual framework to inform progress in simulation research. *Simul Healthc*. 2010;5(1):8-15.
32. Oxman A, Guyatt G. *Users' Guides Interactive: When to Believe a Subgroup Analysis.* 2002. http://ugi.usersguides.org/usersguides/hg/hh_start.asp. Accessed July 15, 2011.