The American Journal of Surgery 220 (2020) 1557–1565

Contents lists available at ScienceDirect

The American Journal of Surgery

journal homepage: www.americanjournalofsurgery.com

Improving ABSITE scores - A meta-analysis of reported remediation models

Tracy J. Cheun [a, b], Mark G. Davies [a, b, *]

[a] Division of Vascular and Endovascular Surgery, Department of Surgery, Long School of Medicine, University of Texas Health at San Antonio, San Antonio, TX, USA
[b] South Texas Center for Vascular Care, South Texas Medical Center, San Antonio, TX, USA

ARTICLE INFO

Article history:
Received 18 May 2018
Received in revised form
20 April 2020
Accepted 20 April 2020

Presented at the Association for Surgical Education (38th Annual Meeting in Austin TX; May 1–3 2018)

Keywords:
ABSITE
Remediation
Meta-analysis

ABSTRACT

Background: The ABSITE is an annual formative assessment of residents' knowledge. This study examines the effects of remediation models on performance in the ABSITE.
Methods: A systemic literature review, qualitative content analysis and a quantitative meta-analysis were performed on studies from 1980 to 2018. Study quality and bias was also assessed. Main outcome measures were extracted to calculate effect sizes using a random effect model.
Results: Seventy-one percent of the studies considered to have acceptable quality and 79% were considered to have a low risk of bias. On qualitative content analysis, the interventions grouped into the following themes: mandatory multimodality remediation program, structured reading program, establishing a passing benchmark, problem-based learning, mandatory didactic conference attendance, learning management system and/or social media, and self-directed learning. Remediation models with the most positive effects were mandatory multimodality remediation programs (SMD 0.78, 95% confidence interval [0.27−1.28] $p = 0.003$) and the use of learning management systems/social media (0.74, [0.32−1.16] $p = 0.001$).
Conclusion: Establishment of mandatory multimodality remediation programs and the use of a learning management systems/social media appear to be the most effective measures.

Introduction

The American Board of Surgery In-training Examination (ABSITE) was first administered in 1975 as an annual formative assessment of surgical residents to track their progress and to prepare for board examinations after completion of residency. It has undergone several iterations and revisions over time and currently is offered as a computer based, multiple choice formative assessment of applied basic science and clinical knowledge in the field of surgery, administered annually to all post-graduate year (PGY) surgical residents in late January. Success on ABSITE has been correlated with success on the American Board of Surgery Qualifying (QE) and Certifying (CE) examinations,[1] and conversely residents who struggle on this exam are at higher risk of failing these board examinations.[2]

A "resident in difficulty" is defined as "a trainee who demonstrates a significant enough problem that requires intervention by someone of authority, usually the program director or chief resident,"[3] and typically struggles with insufficient fund of medical knowledge, poor clinical judgment, inefficient use of time, and/or behavioral issues, any of which can contribute to poor clinical performance and inadequate in -service training examination outcomes. While not endorsed as a sole determinant of resident retention or termination, the ABSITE can help a program identify the "resident in difficulty". Furthermore, in a survey of fellowship program directors, ABSITE performance was the third most important factor in a resident's candidacy for promotion, and 15% of fellowship program directors ranked ABSITE scores as the most important factor in their decision to rank a candidate in the match.[4] Therefore, poor ABSITE performance can ensnare a resident in a self-perpetuating wheel of failure.

Remediation in medical education is an essential part of medical education and may be defined as 'the act of facilitating a correction for trainees who started out on the journey toward becoming a physician but have moved off course'.[5] With the current emphasis on ABSITE performance among program leaderships, there is

* Corresponding author. MHPE Division of Vascular and Endovascular Surgery, South Texas Center for Vascular Care, University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Drive, MC7741, San Antonio, TX, 78229, USA.
E-mail address: daviesm@uthscsa.edu (M.G. Davies).

https://doi.org/10.1016/j.amjsurg.2020.04.028
0002-9610/

obviously a strong interest among surgery residency programs in remediating those residents considered to be "residents in difficulty" and/or residents at risk, and there is a substantial body of work examining the different methods of ABSITE remediation.[6] A systematic review of ABSITE remediation interventions was performed by Kim et al. and the authors concluded that structured reading programs and remedial programs yield consistently beneficial effects on ABSITE performance.[6] However, a quantitative analysis of this body of literature has not been performed. This study is a meta-analysis of the current state of the literature with regard to methods of improving ABSITE performance.

Methods

Study question

Which model or models of remediation have a successful outcome as measured by performance improvement in the ABSITE?

Study identification and inclusion

A systemic literature review was performed from 1980 to 2017 in accordance with PRISMA guidelines. . A search strategy involving three search engines (Ovid, PubMed, and Google Scholar) and six literature databases (MEDLINE, CINAHL- Cumulative Index to Nursing and Allied Health Literature, ERIC-Education Resources Information Center, EBMR - Evidence-Based Medicine Reviews, PsychINFO and EMBASE) for relevant articles was performed with searches being limited to the English language and, in the case of Google Scholar, to the first 20 chronologically listed pages (~200 hits). The date range of the search was 01/01/1980 to 12/31/2018 and was performed in January 2019. Eight search terms and four inclusion criteria were used (Table 1). A broad Boolean search methodology was used Inclusion criteria were studies that quantified the association of a program- or resident-linked variable with ABSITE performance and/or an educational intervention to improve ABSITE outcomes for resident surgeons compared with a no-intervention or a pre-intervention assessment. Abstracts of all identified papers were screened in parallel. Studies after the implementation of the ACGME competencies in 1999/2000 were chosen for inclusion in the meta-analysis to control for different instructional and assessment directives.

Study quality

The quality of studies was assessed using the Medical Education Research Study Quality Instrument (MERSQI), a validated tool designed to objectively evaluate the quality of experimental and observational studies in medical education research.[7] The MERSQI scores study quality based on 6 aspects: study design, sampling, type of data, validity of evaluation instruments, data analysis and outcomes. Each domain offers a maximum score of 3 with an overall maximum score of 18 and minimum score of 5. As a threshold for high- or low-quality studies, the median of the MERSQI scores was used.[8] Studies were also broken down by quartile to further delineate the distribution of the MERSQI scores. Inter-rater reliability was quantified by calculating kappa and intraclass coefficients for each element of the Medical Education Research Study Quality Instrument. Reviewers reviewed all papers and the analysis was random effects, two way, agreement model. The overall The overall median intraclass coefficients was 0.83 (IQR; 0.62, 0.94) with a Cohen Kappa of 0.87.

Risk of bias

The quality of the methodology for each study was assessed using the Newcastle-Ottawa score (NOS).[9] The NOS evaluates three quality parameters (selection, comparability, and outcome) divided across eight specific items, which slightly differ when scoring case control and longitudinal studies Each item on the scale is scored from one point, except for comparability, which can be adapted to the specific topic of interest to score up to two points. As a result, the maximum score for each study is 9, with studies having a score less than 5 points being identified as representing studies with a high risk of bias. Inter-rater reliability was quantified by calculating kappa and intraclass coefficients for the Newcastle-Ottawa score. Reviewers reviewed all papers and the analysis was random effects, two way, agreement model. The overall median intraclass coefficients was 0.89 (IQR; 0.81, 0.94) with a Cohen Kappa of 0,91.

Qualitative content analysis

A qualitative directed content analysis was used to analyze the studied interventions and organize them into conceptual interventional categories. The interventional categories were informed by the classification of Kim et al..[6] To conduct the analysis in a constant comparative manner, two reviewers (MGD and TJC) analyzed and identified themes of remediation models. Following review of the initial themes, the final themes and subthemes were then discussed and agreed upon by consensus.

Quantitative

Meta-analysis was conducted in accordance with the Institute of

Table 1
Search strategy.

In training Examination
In Service Examination
In service
ABSITE
American Board of Surgery
Remediation
Assessment
Residency
MeSH terms
Education, Medical
Graduate/methods
Educational Measurement
General Surgery/education
Humans
Internship and Residency
Specialty Boards, United States

Medicine (IOM) standards.[10] We conducted a meta-analysis for intervention versus no intervention or pre-intervention assessment for all seven identified themes of remediation models. Main outcome measures of the studies were reported as raw score (percent correct), percentile score, and percentile rank. A Hedge's 'g' effect size and a Cohen 'd' statistic were calculated as previously described, when the mean and standard deviation or odds ratio of two groups were present and convertible to a standardized mean difference.[11–13] In cases where these variables were not reported, the effect size was estimated using the outcomes of statistical test results (e.g. P values).[11] For studies with 2-group pre-test, post-test studies, post-test means were adjusted for pre-test or adjusted statistical test results. If these were not available, the difference in change scores using the pre-test variance was standardized.[12] In reports with a crossover design, the means or exact statistical test results adjusted for repeated measures were used if these data points were not available, the means pooled across each intervention were used.[14,15] For those reports that did not report either a P-values nor any measure of variance, the average standard deviation from all other studies included in the analysis was used. Interpretations of educational significance utilized Cohen's effect size classifications for standard mean differences (SMD), <0.2 was defined as small, 0.21–0.50 as medium, and >0.51as large.[16] The SMD of ABSITE improvement compared to no intervention, along with 95% confidence interval, was used as the summary statistic for each theme. The weight and effect size of each study effect was determined by calculating the within-study variance. We pooled effect sizes using random effects model. Total between-study variance was determined where significant heterogeneity existed. Heterogeneity was assessed using an I^2 calculation. Statistical processing was performed on MedCalc (MedCalc Software, Ostend, Belgium).

Results

Study description

Study flow and selection are demonstrated in the PRISMA diagram in Fig. 1. Of 98 articles identified using the described search strategy, 34 met the defined criteria for analysis. Across these studies, 2847 residents were included encompassing all PGY levels of training in general surgery, and study sample sizes ranged from 8 to 408 residents (Table 2). On review of the studies, there were three reported outcome measures and three study designs (Table 2). Outcome measures were: percentile ABSITE score (40% of studies), raw ABSITE score (38% of studies) and ABSITE percentile rank (17%). Study designs were categorized as: comparison of pre- and post-intervention ABSITE performance (68%), observational (18%), and comparison of control versus study group (15%). Pre- and post-intervention typically observed exam performance from one year to the next after implementation of a specific programmatic intervention such as a structured reading program (Table 2). Observational studies examined correlations between various factors, such as conference attendance or previous exam performance, and ABSITE performance. Controlled studies typically compared the implementation of an intervention, for example attendance at problem-based learning sessions, on one cohort of residents against another cohort following the residency's traditional study program.

Study quality

The MERSQI scores ranged between 9 and 14.5 with an overall mean of 11.4 ± 2.6 (mean ± SD). A breakdown of overall performance by domain is detailed in Table 3. The intraclass correlation co-efficient for each domain of the MERSQI were acceptable (Table 3). Collectively, nearly all studies performed very well in the type of data collected, the type of data analysis used and originality. In terms of study design, sampling and outcomes, the studies in general were adequate. Validity of the evaluation instrument used was weak as the majority of studies lacked any kind of evaluation instrument, and those that did were limited to surveys (Table 3). The median MERSQI score was 11.0 (IQR, 10.5, 12); and using this threshold as described by Cook et al.,[8] 71% of the studies had acceptable quality and 29% were considered poor quality. However, if one segments the MERSQI scores by quartiles only 36% were in the upper quartiles suggesting superior quality and the remainder in the lower quartiles suggesting lower quality (21%, placed high in the first quartile; 15% placed in the second quartile; 36% placed in the third quartile and 26% were placed in the fourth quartile). When the risk of bias was assessed using the Newcastle Ottawa Score the median score was 11 (IQR, 9,13); 79% has scores> 5 and were considered to have a low risk of bias, while 21% had score of 5 or less and were considered to have a high risk of bias (Table 2).

Qualitative analysis

Using a qualitative directed content analysis and two reviewers, we identified 29 unique codes among the 34 studies included in the qualitative synthesis. These were distilled into seven domains: mandatory multimodality remediation program, structured reading program, establishing a passing benchmark, problem-based learning, mandatory didactic conference attendance, learning management system and/or social media, and self-directed learning (Table 4). These categories largely correlated with those outlined by Kim et al. in their systematic review with the additional distinction of established passing benchmarks and learning management systems.[6]

Quantitative analysis

Of the 34 studies, 38% had a "large" effect size. 6% had a medium effect size and the remainder had a "small" effect size based on Hedge's 'g' and Cohen "d" statistics (Tables 4 and 5). Taking the entire set of studies together, the Standard Mean Difference (SMD) was 0.52 (0.39–0.65 95th CI, $p < 0.001$) suggesting an overall medium effect of remediation (Fig. 2). The heterogeneity of the data was high (Cochran's Q, 242.0, $P < 0.0001$) with the I^2 for the data set, which describes the percentage of total variation across studies that is due to heterogeneity rather than chance, was 80.6% (74.8–85.0, 95th CI). The studies were then grouped into the seven themes derived from the qualitative directed content analysis (Fig. 3). Remediation models with a large effect were: mandatory multimodality remediation programs (SMD for improvement, 0.78, 95% confidence interval [0.27–1.28]; $P = 0.003$) and the use of learning management systems/social media (0.74, [0.32–1.16]; $P = 0.001$). Remediation models with lesser effect were mandatory didactic conference attendance (0.52, [0.39–0.65]; $p = 0.001$) structured reading programs (0.52, [0.30–0.73]; $P = 0.001$) and self-directed learning (0.51, [0.14–0.87]; $P = 0.006$). Remediation models with small effects were problem-based learning (0.45, [0.16–0.74]; $P = 0.003$) and establishing a passing benchmark (0.29 [0.04–0.54]; $P = 0.02$).

Discussion

There are four core components of a successful remediation program[1]: initial assessment (or screening) using multiple assessment tools to identify deficiencies,[2] diagnosis of problems and development of an individualized learning plan,[3] provision of

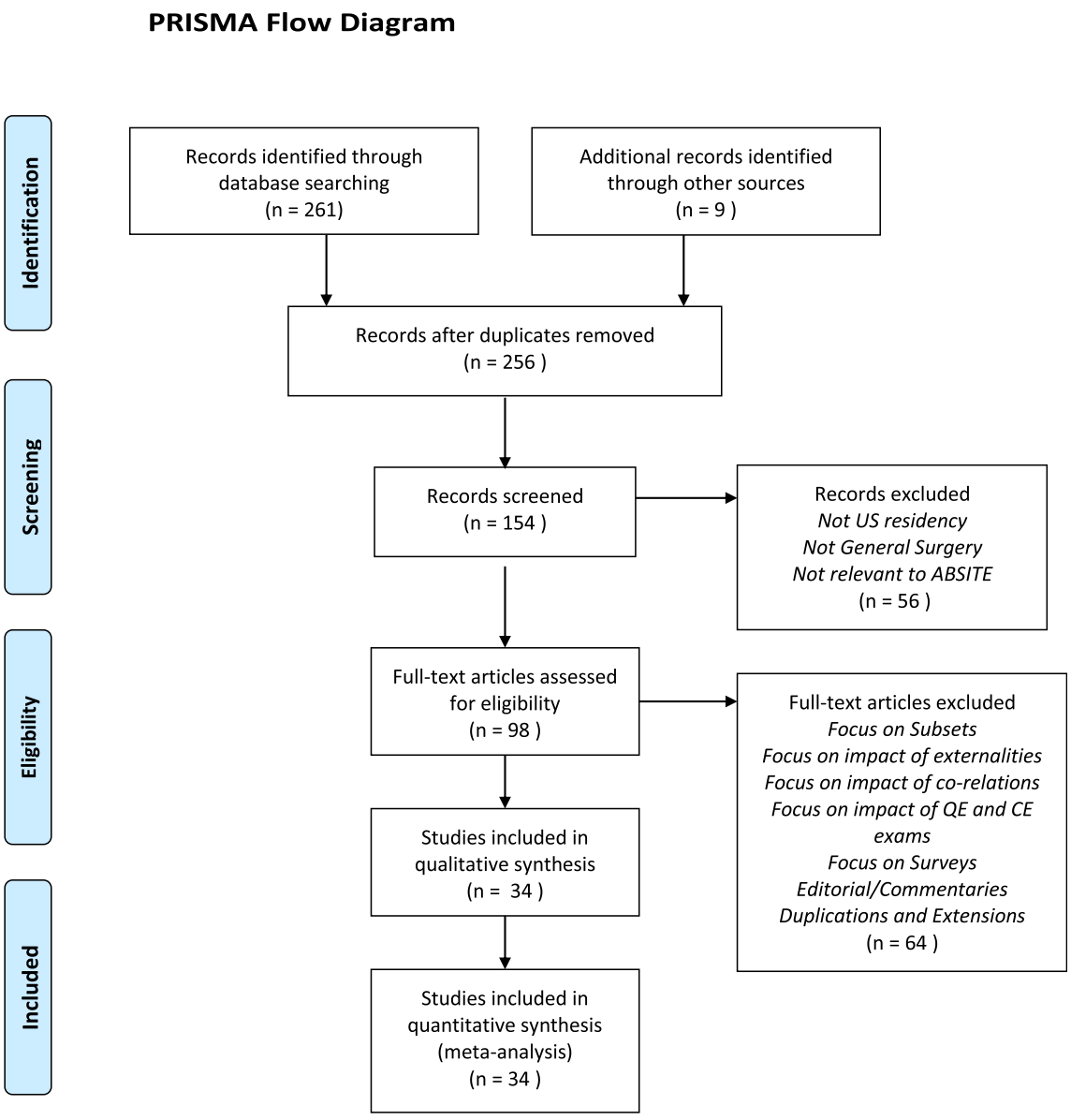

Fig. 1. Prisma flow diagram.

instruction that includes deliberate practice, feedback, and reflection, and[4] reassessment and certification of competence.[17] The ABSITE is currently used as a screening tool to identify deficiencies. Multiple drivers have been previously associated with the success or failure of surgical residents on the ABSITE.[6] Most of the literature has concentrated on pre-existing factors to preemptively screen and identify the 'at risk; candidate. However, there remains the problem of the resident who is having difficulty and how to remediate their performance early so that they can be successful in the post-graduation professional exams of the ABS. A broad range of remediation models have been the subject of research reports in recent years covering a wide range of methods. This study is the first meta-analysis of models of ABSITE remediation in the literature. The results identify significant improvement in ABSITE performance is associated with establishment of mandatory multimodality remediation programs and the use of a learning management systems/social media environment.

Establishment of mandatory multimodality remediation programs appears to be successful but the components needed remains to be fully defined. De Virgilio showed that an educational program of weekly assigned reading followed by weekly examinations prepared and administered by the program director, can resulted in an increase in the 5-year first-time pass rates on the ABS Qualifying and Certifying Exams and combined QE/CE.[18] Harthun et al. reported on a faculty-driven program wherein a faculty member was designated to meet with each resident with below standard ABSITE scores released in March. In April, individual study plans were designed with each resident which emphasized practice questions and strategic planning for study incorporating resources such as the Surgical Education and Self-Assessment Program (SESAP) review. ABSITE test structure and results were reviewed as well. These individualized programs minimized the stigma of poor test results, controlled the time commitment required by the residents, and maximized the benefits of a question-based study system. In addition, acquisition of ABSITE exam-taking skills resulted in significant improvements in scores the following year. However, maintenance of these results required perennial efforts and commitments by select faculty and continued evaluation by the

Table 2
Study characteristics.

Number	Year	Citation	Reference Number	Study period	N	PGY levels	intervention type	Study Design	Outcome Variable	Newcastle Ottawa Score
1	1982	Shetler	[37]	1975 –1981	13	1,2,3,4,5	Attendance at weekly teaching conference	Observation	Percentile Rank	6.5
2	1984	Dean et al.	[38]	1981 –1983	44	1,2,3,4,5	Structured reading program	Pre- and post-intervention comparison	Percentile Rank	11.5
3	1988	Pollak and Baker	[39]	1986 –1987	169	1,2,3,4,5	Attendance at weekly teaching conference	Observation	Standard Score	4.5
4	1991	Hirvela and Becker	[40]	1988 –1990	84	1,2,3,4,5	Structured reading program	Pre- and post-intervention comparison	Standard Score	9.5
5	1992	Safran et al.	[41]	1989 –1991	106	1,2,3,4,5	Attendance at weekly teaching conference Mock Orals	Pre- and post-intervention comparison	Standard Score	10
6	1997	Itani et al.	[42]	1995 –1996	64	1,2,3,4,5	Attendance at weekly teaching conference Structured lecture program	Observation	Standard Scores	5
7	1999	Pemberton et al.	[43]	1995 –1998	10	1,2,3,4	Mandatory remedial program Weekly ABSITE-styled test questions ABSITE-styled practice exams	Pre- and post-intervention comparison	Percentile Score Standard Scores	11
8	2002	Pofahl et al.	[36]	1991 –2000	76	1,2,3,4,5	Expected standard of 35th percentile	Pre- and post-intervention comparison	Percentile Rank	10
9	2003	de Virgilio et al.	[44]	1999 –2001	25	1,2,3,4,5	Structured reading program Weekly ABSITE-styled test questions	Pre- and post-intervention comparison	Percentile Rank	13
10	2005	de Virgilio et al.	[45]	1999 –2004	19	1,2,3,4,5	Structured reading program Weekly ABSITE-styled test questions	Pre- and post-intervention comparison	Percentile Rank	13
11	2005	Harthun et al.	[19]	2000 –2004	8	1,2,3,4,5	Mandatory remedial program	Observation	Standard Scores Percentile Rank	6
12	2006	Nguyen et al.	[28]	2004 –2006	55	1,2,3,4,5	Structured reading program Problem based Learning groups	Pre- and post-intervention comparison		13
13	2006	Ahmed et al.	[46]	2002 –2005	49	1,2,3,4,5	Attendance at weekly teaching conference	Pre- and post-intervention comparison	Percentile Rank	12
14	2006	Mahmoud et al.	[47]	1997 –2002	31	1,2,3,4,5	Attendance at weekly teaching conference	Pre- and post-intervention comparison	Percentile Rank Standard Scores	13
15	2006	Mahmoud et al.	[47]	1997 –2002	31	1,2,3,4,5	Attendance at weekly teaching conference	Pre- and post-intervention comparison	Percentile Rank Standard Scores Correct Scores	13
16	2006	Borman	[48]	2003 –2004	15	1,2,3,4,5	Mandatory remedial program	Control vs study group	Percentile Rank Standard Scores	9
17	2006	Ferguson and Warshaw	[24]	2004 –2005	19	1,2,3,4,5	Voluntary access to web-based educational tool (BeST resident)	Pre- and post-intervention comparison	Standard Scores	12
18	2008	de Virgilio et al.	[18]	1997 –2007	49	1,2,3,4,5	Structured reading program Weekly ABSITE-styled test questions	Pre- and post-intervention comparison	Percentile Rank	13
19	2008	Gregg et al.	[49]	2005 –2007	25	1,2,3,4	Structured reading program Weekly ABSITE-styled test questions Problem based Learning groups	Pre- and post-intervention comparison	Standard Score	13
20	2008	Farrohki et al.	[50]	2004 –2006	75	1,2,3,4,5	Attendance at weekly teaching conference	Pre- and post-intervention comparison	Percentile Rank	13
21	2008	Lee et al.	[29]	2004 –2006	42	1,2,3,4,5	Problem based Learning groups	Pre- and post-intervention comparison	Percentile Rank	13
22	2008	Kosir et al.	[51]	2007 –2008	31	2,3,4,5	Mandatory remedial program	Control vs study group	Percentile Rank Correct Scores	11
23	2010	Lube et al.	[52]	1994 –2009	208	1,2,3,4,5	Structured lecture program Bimonthly ABSITE-styled test questions	Pre- and post-intervention comparison	Standard Score Percentile Rank	12
24	2011	Corneille et al.	[53]	2005 –2010	263	1,2,3,4,5	ABSITE-styled practice exams	Pre- and post-intervention comparison	Percentile Score	11
25	2013	Willis et al.	[54]	2011 –2012	20	1	Fundamental of surgery curriculum	Control vs study group	Percentile Rank	11
26	2013	Krajewski et al.	[55]	2011	108	1	Boot Camp	Pre- and post-intervention comparison	Percentile Rank	9
27	2014	Dua et al.	[56]	2008 –2009	508	3,4	A multimedia review course within an online LMS	Pre- and post-intervention comparison	Percentile Score	8

(*continued on next page*)

Table 2 (*continued*)

Number	Year	Citation	Reference Number	Study period	N	PGY levels	intervention type	Study Design	Outcome Variable	Newcastle Ottawa Score
							ABSITE-styled practice exams Structured reading program			
28	2015	Buckley et al.	[57]	2010 –2014	77	1,2,3,4,5	Structured reading program Weekly ABSITE-styled test questions Mock Orals	Control vs study group	Percentile Score	11
29	2015	Kelly et al.	[58]	2000 –2009	140	1	A multimedia review course within an online LMS	Observation	Standard Scores	7
30	2016	Willis et al.	[59]	not defined	206	1,2,3,4,5	Question Writing	Control vs study group	Percentile Scores	14
31	2017	Lamb et al.	[60]	2016 –2017	46	1,2,3,4,5	ABSITE-styled practice exams Microblogging	Pre- and post-intervention comparison	Percentile Rank	8
32	2018	Decoteau et al.	[20]	2001 –2016	52	1,2,3,4,5	Structured reading program Mandatory remedial program Problem based Learning groups	Pre- and post-intervention comparison	Percentile Score	11
33	2018	Tarabichi et al.	[61]	2015 –2017	45	1,2,3,4,5	ABSITE-styled practice exams	Pre- and post-intervention comparison	Percentile Score	12
34	2018	Kanter et al.	[62]	2013 –2016	134	1,2,3,4,5	ABSITE-styled practice exams	Observation	Percentile Score Correct Scores	7

program.[19] Decoteau et al. demonstrated significant improvements in ABSITE performance as well as ABS Qualifying Exam performance using a multimodal approach consisting of a passing benchmark at the 30th percentile, problem-based learning sessions, didactic conferences, use of a learning management system, and focused remediation programs paired with mentorship and identification of individual weaknesses.[20] Visconti et al.[21] demonstrated that individualized education plans (IEPs) for emergency medicine residents, combining self-study audio review lectures with short-answer examinations has been shown to significantly improve board pass rates. Similarly, Mathis et al.[22] in an internal medicine residency program designed a specific, multiple-choice testing program and a separate board review program, both administered during a continuous long-block elective experience during the twelve months between the second post-graduate year (PGY-2) and PGY-3 in-training examinations and demonstrated that multimodality approach resulted in a significant increase in median individual IM-ITE percentile score between PGY-2 and PGY-3 examinations (8.5 vs. 1.0 percentile point increase, intervention vs. control). The current findings suggest that a multi-modality program can be successful in remediation performance on a formative assessment.

There is an increasing use of LMS and/or social media systems to provide asynchronous education across all medical educations specialties. Tarras et al.[23] have documented that an increased usage of SCORE portal was associated with higher performance on the ABSITE. In a study where all PGY1 and PGY2 residents were given unlimited, self-directed access to a web-based educational tool that covered the basic science of surgery and basic concepts of clinical surgery through readings, problem-based learning, case-based learning, and practice tests, Ferguson et al. were unable to demonstrate a significant difference in ABSITE scores before or after the use of the Web-based educational tool.[24] There was no significant relationship between use of the tool (either in total time or total tutorials completed) and ABSITE score. For PGY2 residents, there was a negative relationship between total time spent on the program and ABSITE score as well as between total tutorials completed and ABSITE score. This failure to demonstrate a benefit was attributed in part to the voluntary nature of the LMS use, and the authors suggested that mandatory implementation may change their results. This concept of a mandated use of the LMS was shown in the study by Decoteau et al.[20] A later study by Dua et al. used a structured surgeon-directed LMS to provide a multimedia-oriented review course of print and digital media associated with practice questions, review textbooks, weekly reading assignments, and slide and audio reviews by PGY3 and PGY4 residents. The results of the Dua study demonstrated that a consistent multi-year structured approach framed by the LMS platform led to enhanced performance on the ABSITE with scores higher the longer the resident used the course on the LMS.[25] Drake et al.[26] demonstrated that a web-based directed reading program which matched internal medicine residents to reading assignments based on their individual ITE-failed educational objectives and provides direct electronic feedback from their teaching physicians was successful linked to improved performances on subsequent exams. To that point, more studies on the effects of online learning management systems and social media platforms are needed. In the current age of large multi-center training programs, where residents can be stretched across a dozen hospitals in multiple cities and working at different times of the day, there is clear utility in supplementing an

Table 3
MERSQI domain and item scores, intraclass correlation coefficients.

	Mean	SD	Median	Percent of maximal domain score	intraclass correlation co-efficients
OVERALL	11.3	1.3	11.0		
DOMAIN					
Study design	1.5	0.4	1.5	50%	0,83
Sampling	2.0	0.5	2.0	67%	0.74
Type of data	2.9	0.5	3.0	98%	0.91
Validity of evaluation instruments	0.4	0.8	0.0	15%	0,71
Data analysis	2.6	0.7	3.0	86%	0.79
Outcomes	1.5	0.3	1.5	49%	0.76
Originality	0.9	0.2	1.0	90%	0.99

Table 4
Distribution of studies in each Effect size, the associated MERSQI score, and number of Quality Studies (Score ≥11).

Effect size	Studies (n)	%	MERSQI Score		Quality Studies (n)
			Mean	SD	
Large	13	38%	11.5	1.0	4
Medium	2	6%	11.4	1.7	3
Small	17	50%	11.2	1.3	4

ABSITE study program with a web-based platform that is accessible at all times from any location.

In a recent study, gamification of Twitter was implemented and observed for effect on ABSITE performance. Participation was voluntary and consisted of daily incentivized participation in open-ended questions and discussions that coordinated with the following week's didactic conference topic. Participants significantly increased ABSITE scores from 2016 to 2017, while non-participants had significantly decreased scores.[27] A follow-up study showed that non-participants cited hesitation to download and utilize twitter in the face of existing time constraints and mobile device limitations. The study was limited in being underpowered as well as susceptible to self-selection. However, the relatively low cost of implementation of social media in surgical education makes this an easily explorable modality whose role has yet to be characterized.

Problem-based learning models have been incorporated in the Undergraduate Medical Education curriculum for several years. In Graduate Medical Education, problem-based learning models have a mixed performance in the literature given that they are rarely used or added to a program in isolation. One study by Nguyen et al. observed effects of a combined structured reading program with faculty-led weekly problem-based learning sessions. The structured reading program was associated with improvements in ABSITE performance from pre- and post-intervention, however the problem-based learning groups were assessed only by surveying the pilot year participating residents. While it was viewed positively overall by residents, it alone was not associated with any improvement in ABSITE performance despite the success of the structured reading program.[28] Lee et al. reported ABSITE score improvement associated with problem-based learning class attendance, however this trend did not reach statistical

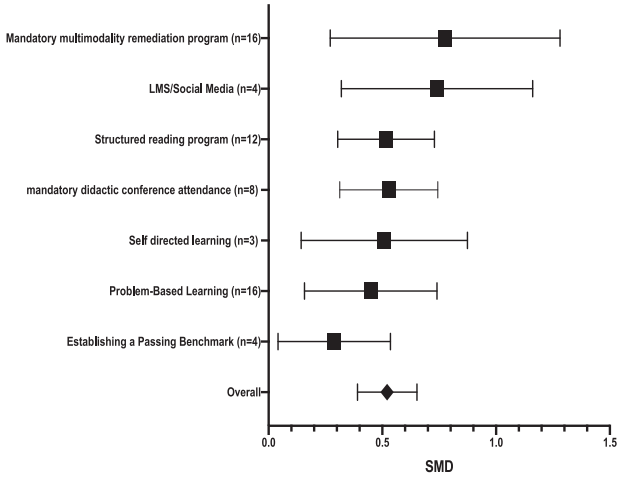

Fig. 3. Modified Forest plot of aggregate score for each thematic intervention with changes expressed as a standard mean difference (SMD) and 95% of confidence interval for each thematic intervention. ■ = the pooled effect size for each theme; horizontal line = 95% of confidence interval; ◆ = pooled effect size for all themes.

significance, nor did it account for self-selection (more motivated residents have better attendance).[29] In addition, they attribute much of their success to the resident-faculty interactions that the PBL groups created. This perhaps suggests that the true benefit lies more in increasing small-group contact between residents and faculty rather than in the PBLs themselves. A controlled observational stud in pediatric residency has shown no difference in standardized exam performance associated with a problem-based learning curriculum.[30] Problem-based learning hinges upon small groups, 8 or fewer typically, with questions and discussions of clinically-based scenarios. While these may be viewed positively by residents and contribute to their critical thinking skills and application of basic science and clinical skills, their effectiveness in making them better exam takers is questionable. Particularly in the age of the 80-hr work week, where efficiency and efficacy are paramount to a successful study regimen, PBL is typically not an efficient method of covering the massive amount of material that the ABSITE covers. This is not to say that PBL does not have a role in

	Control Group			Intervention Group			Std Mean Differnce (95%CI)		
Study	Mean	SD	N	Mean	SD	N	SMD	Lower 95% CI	Upper 95% CI
1982, Shetler	69	7	101	72	6	92	-0.5	-0.7	-0.2
1984, Dean et al	46	47	47	71	47	49	-0.5	-0.9	-0.1
1988, Pollak and Baker	52	17	47	69	19	159	-0.9	-1.2	-0.6
1991, Hirvela and Becker	146	50	47	171	49	49	-0.5	-0.9	-0.1
1992, Safran et al.	52	19	14	67	7	18	-1.1	-1.8	-0.3
1997, Itani et al	49	23	213	60	18	55	-0.5	-0.8	-0.2
1999, Pemberton et al	49	23	213	60	18	55	-0.5	-0.8	-0.2
2002, Pofahl et al	36	28	107	43	22	128	-0.3	-0.5	0.0
2003, de Virgilio et al	29	29	33	44	31	11	-0.5	-1.2	0.2
2005, de Virgilio et al	69	7	101	72	6	92	-0.5	-0.7	-0.2
2005, Harthun et al.	52	17	1	69	19	159	-0.9	-1.2	-0.7
2006, Nguyen et al	52	19	14	67	7	18	-1.1	-1.8	-0.3
2006, Ahmed et al.	438	72	55	542	107	85	-1.1	-1.5	-0.7
2006, Mahmoud et al.	54	5	27	61	4	27	-1.5	-2.1	-0.9
2006, Mahmoud et al.	52	17	508	69	19	159	-1.0	-1.2	-0.8
2006, Borman	30	44	7	88	70	8	-0.9	-2.0	0.1
2006, Ferguson and Warshaw	532	57	19	530	46	19	0.0	-0.6	0.7
2006, de Virgilio et al.	30	44	7	88	70	8	-0.9	-2.0	0.1
2008, Gregg et al	65	47	25	71	47	25	-0.1	-0.7	0.4
2008, Farrohki et al	29	29	33	44	31	11	-0.5	-1.2	0.2
2008, Lee et al	54	5	27	61	4	27	-1.5	-2.1	-0.9
2008, Kosir et al	0	8	11	4	8	18	-0.5	-1.2	0.3
2008, Lube et al	69	7	101	72	6	92	-0.5	-0.7	-0.2
2008, Corneille et al	52	19	14	67	7	18	-1.1	-1.8	-0.3
2010, Krajewski et al	46	28	107	53	22	128	-0.3	-0.5	0.0
2011, Willis et al.	439	121	10	485	107	10	-0.4	-1.3	0.5
2013, Dua et al	52	17	33	69	19	11	-1.0	-1.7	-0.2
2013, Buckley et al	52	19	14	67	7	18	-1.1	-1.8	-0.3
2014, Kelly et al	55	10	55	66	9	85	-1.2	-1.5	-0.8
2015, Willis et al.	50	21	107	54	20	99	-0.2	-0.5	0.1
2015, Lamb et al	50	21	107	54	20	99	-0.2	-0.5	0.1
2016, Decoteau et al	52	19	14	67	7	18	-1.1	-1.8	-0.3
2017, Tarabichi et al	72	7	45	71	7	45	0.1	-0.3	0.6
2018, Kanter et al	73	7	40	66	10	30	0.8	0.3	1.3
Total (95%CI)			2304			1925	-0.599	-0.781	-0.447

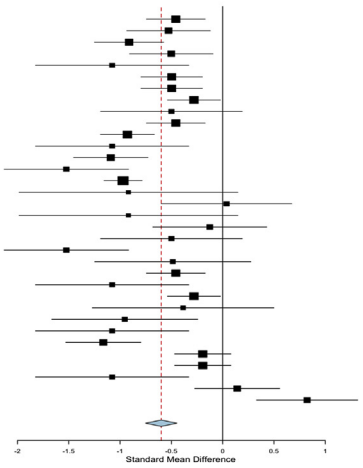

Fig. 2. Forest plot of ABSITE score for each study changes expressed as a standard mean difference (SMD) and 95% of confidence interval of the intervention for each of the studies. ■ = the effect size for one trial; horizontal line = 95% of confidence interval; ◆ = pooled effect size for all trials.

a surgical residency curriculum. However, its use as a means of improving ABSITE performance may be misdirected especially when it comes to the "resident in difficulty" who may lack the confidence to actively participate and has every opportunity to be a wallflower during PBL sessions.

In the current study both structured reading programs and mandatory conference attendance had an intermediate effect on performance. In the review by Kim et al.,[6] the authors identified multiple reports from general surgical residencies that demonstrated that structured reading programs can improve ABSITE performance. In fact, all studies of structured reading programs showed improvement in ABSITE performance. Our data suggests that this in isolation is not the best use of resources. Early reports suggested that mandatory conference attendance can improve ABSITE performance,[6] Our analysis suggests that mandatory conference attendance is not as effective as early reports suggested. Gene Hern et al.[31] demonstrated that greater conference attendance by emergency medicine residents does not correlate with performance on an individual's ITE scores Cacamese et al.[32] also explored the relationship between attendance at conferences during residency training and residents' performance on the In-Training Examination (ITE) in Internal Medicine and showed that there was no correlation between prior conference attendance and ITE scores. McDonald et al.[33] disputed this finding suggesting that attendance at a core curriculum conference rather than grand rounds or mortality and morbidity did influence the In-Training Examination (ITE) in Internal Medicine.

In 2014, Simpson-Camp et al. reported that general surgery residents were unable to accurately predict their ABSITE performance immediately prior to, or even after taking the examination. In general, residents overestimated their performance.[34] The authors suggested that the overestimation reflected a poor understanding by the residents of what one should consider adequate and inadequate preparation for the examination. In a survey of surgery program directors, 69% of respondents stated that 30th percentile was used as the established passing benchmark, followed by 35th and 40th percentiles.[35] Establishing a performance criterion has been shown to improved ABSITE scores[36] After institution of a passing benchmark of 35th percentile or higher, the proportion of "failing" scores decreased significantly, and the proportion of performances at or above the national average increased significantly as well. Decoteau et al.[20] demonstrated that the value of incorporating a passing benchmark at the 30th percentile in the success of a multimodality remediation program. Given Simpson-Camp's findings, these expectations must be communicated to residents well in advance of the exam along with a clear plan to remediate those that fail to meet the benchmark in order to be effective. Use of a benchmark can be considered a useful screening tool to channel a resident into a remediation program.

Limitations

Limitations of this meta-analysis lie primarily in the heterogeneity of the studies included, from variable outcome measures reported to omission of sample size. More studies conducted with a more vigorous statistical analysis would improve the strength of the meta-analysis. This methodological problem is in part due to the retrospective observational nature of most of the studies as well as to the fact that data were derived from final publications, not from the authors' raw datasets. In addition, many of the studies' sample sizes relied on voluntary resident participation, creating room for self-selection bias. As in any literature review, an exhaustive search of the published literature inevitably omits the unpublished works which are often characterized by a lack of statistically significant findings. Additionally, there were few studies that examined the effects of LMS and social media which are burgeoning tools in the medical education armamentarium, and likely the effects of these methods are thus far understated in the published literature.

Conclusion

Using a meta-analysis strategy, it appears to effectively assist residents with poor ABSITE scores, establishment of mandatory multimodality remediation programs and the use of learning management systems/social media appear to be the most effective measures that can result in improved ABSITE performance. Further studies in the field should examine the effectiveness of a synthesis of these methods as well as the appropriate role of social medial and online platforms.

Contributions

Tracy Cheun: concept and design, performance of the work, analysis and/or interpretation of data; critical writing or revising the intellectual content; and final approval of the version to be published.

Mark G. Davies: concept and design, analysis and/or interpretation of data; critical writing or revising the intellectual content; and final approval of the version to be published.

Funding

This research received no specific grant support from any funding agency in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.amjsurg.2020.04.028.

References

1. Jones AT, Biester TW, Buyske J, et al. Using the American Board of Surgery In-Training Examination to predict board certification: a cautionary study. *J Surg Educ*. 2014;71(6):e144–e148.
2. De Virgilio C, Yaghoubian A, Kaji A, et al. Predicting performance on the American Board of Surgery qualifying and certifying examinations: a multi-institutional study. *Arch Surg*. 2010;145(9):852–856.
3. Dupras DM, Edson RS, Halvorsen AJ, et al. "Problem residents": prevalence, problems and remediation in the era of core competencies. *Am J Med*. 2012;125(4):421–425.
4. Miller AT, Swain GW, Widmar M, Divino CM. How important are American Board of Surgery In-Training Examination scores when applying for fellowships? *J Surg Educ*. 2010;67(3):149–151.
5. Kalet A, Chou CL. *Remediation in Medical Education*. Springer; 2014.
6. Kim RH, Tan T-W. Interventions that affect resident performance on the American Board of Surgery In-Training Examination: a systematic review. *J Surg Educ*. 2015;72(3):418–429.
7. Reed DA, Cook DA, Beckman TJ, et al. Association between funding and quality of published medical education research. *J Am Med Assoc*. 2007;298(9):1002–1009.
8. Cook DA, Hatala R, Brydges R, et al. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. *J Am Med Assoc*. 2011;306(9):978–988.
9. Wells G, Shea B, O'Connell D, et al. *The Newcastle-Ottawa Scale (NOS) for Assessing the Quality of Nonrandomised Studies in Meta-Analysis*. 2011 [website].
10. Institute of Medicine. *Finding what Works in Health Care: Standards for Systematic Reviews*. Washington DC: Institute of Medicine of the National Academies; 2011.
11. Borenstein M. In: Cooper H, Hedges LV, Valentine JC, eds. *Effect Sizes for Continuous Data*. New York, NY: Russell Sage Foundation; 2009.
12. 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.
13. Hunter JE, Schmidt FL. *Methods of Metaanalysis: Correcting Error and Bias in Research Findings*. housand Oaks, CA: Sage Pulications; 2004.
14. 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.
15. Higgins JPT, Green S. *Cochrane Handbook for Systematic Reviews of Intervention*; 2011. Available from: http://www.cochrane.org/resources/handbook/index.htm.
16. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. second ed. Hillsdale, NJ: Lawrence Erlbaum; 1988.
17. Hauer KE, Ciccone A, Henzel TR, et al. Remediation of the deficiencies of physicians across the continuum from medical school to practice: a thematic review of the literature. *Acad Med*. 2009;84(12):1822–1832.
18. de Virgilio C, Chan T, Kaji A, Miller K. Weekly assigned reading and examinations during residency, ABSITE performance, and improved pass rates on the American Board of Surgery Examinations. *J Surg Educ*. 2008;65(6):499–503.
19. Harthun NL, Schirmer BD, Sanfey H. Remediation of low ABSITE scores. *Curr Surg*. 2005;62(5):539–542.
20. Decoteau MA, Louis Rivera L, Umali K, et al. A multimodal approach improves American Board of Surgery In- Training Examination scores. *Am J Surg*. 2018;215:315–321.
21. Visconti A, Gaeta T, Cabezon M, et al. Focused board intervention (FBI): a remediation program for written board preparation and the medical knowledge core competency. *J Grad Med Educ*. 2013;5(3):464–467.
22. Mathis BR, Warm EJ, Schauer DP, et al. A multiple choice testing program coupled with a year-long elective experience is associated with improved performance on the internal medicine in-training examination. *J Gen Intern Med*. 2011;26(11):1253–1257.
23. Tarras S, Edelman DA. SCORE usage predicts remediation needs after ABSITE in surgery residents. *J Am Coll Surg*. 2019;229(4):e126.
24. Ferguson CM, Warshaw AL. Failure of a Web-based educational tool to improve residents' scores on the American Board of Surgery in-training examination. *Arch Surg*. 2006;141(4):414–416.
25. Dua A, Sudan R, Desai SS. Improvement in American Board of Surgery in-training examination performance with a multidisciplinary surgeon-directed integrated learning platform. *J Surg Educ*. 2014;71(5):689–693.
26. Drake SM, Qureshi W, Morse W, Baker-Genaw K. A time-efficient web-based teaching tool to improve medical knowledge and decrease ABIM failure rate in select residents. *Med Teach Online*. 2015;20(1):29221.
27. Lamb LC, DiFiori MM, Jayaraman V, et al. Gamified twitter microblogging to support resident preparation for the american board of surgery In-Service training examination. *J Surg Educ*. 2017;74(6):986–991.
28. Nguyen L, Brunicardi FC, DiBardino DJ, et al. Education of the modern surgical resident: novel approaches to learning in the era of the 80-hour workweek. *World J Surg*. 2006;30(6):1120–1127.
29. Lee L, Brunicardi FC, Scott BG, et al. Impact of a novel education curriculum on surgical training within an academic training program. *J Surg Res*. 2008;145(2):308–312.
30. Ozuah PO, Curtis J, Stein RE. Impact of problem-based learning on residents' self-directed learning. *Arch Pediatr Adolesc Med*. 2001;155(6):669–672.
31. Gene Hern Jr H, Wills C, Alter H, et al. Conference attendance does not correlate with emergency medicine residency in-training examination scores. *Acad Emerg Med*. 2009;16:S63–S66.
32. Cacamese SM, Eubank KJ, Hebert RS, Wright SM. Conference attendance and performance on the in-training examination in internal medicine. *Med Teach*. 2004;26(7):640–644.
33. Associations of conference attendance with internal medicine in-training examination scoresMcDonald FS, Zeger SL, Kolars JC, eds. *Mayo Clin Proc*. 2008;83(4):449–453.
34. Simpson-Camp L, Meister EA, Kavic S. Surgical resident accuracy in predicting their ABSITE score. *JSLS: JJ Sic Lap Surg*. 2014;18(2):277.
35. Kim JJ, Kim DY, Kaji AH, et al. Reading habits of general surgery residents and association with American Board of Surgery in-training examination performance. *JAMA Surg*. 2015;150(9):882–889.
36. Pofahl WE, Swanson MS, Cox SS, et al. Performance standards improve American Board of Surgery in-training examination scores. *Curr Surg*. 2002;59(2):220–222.
37. Shetler PL. Observations on the American Board of Surgery in-training examination, board results, and conference attendance. *Am J Surg*. 1982;144(3):292–294.
38. Dean R, Hanni C, Pyle M, Nicholas W. Influence of programmed textbook review on American board of surgery in-service examination scores. *Am Surg*. 1984;50(7):345–349.
39. Pollak R, Baker R. The acquisition of factual knowledge and the role of the didactic conference in a surgical residency program. *Am Surg*. 1988;54(9):531–534.
40. Hirvela ER, Becker DR. Impact of programmed reading on ABSITE performance. *Am J Surg*. 1991;162(5):487–490.
41. Safran DB, Crombie HD, Allen L, et al. Preliminary assessment of a scientific curriculum in a surgical residency program. *Arch Surg*. 1992;127(5):529–535.
42. Itani KM, Miller CC, Church HM, Mccollum CH. Impact of a problem-based learning conference on surgery residents' in training exam (ABSITE) scores. *J Surg Res*. 1997;70(1):66–68.
43. Pemberton LB, Van Way C, Schroder WB. Special curriculum for surgery residents who fail the ABSITE. *Curr Surg*. 1999;56(3):155–160.
44. de Virgilio C, Stabile BE, Lewis RJ, Brayack C. Significantly improved American Board of Surgery In-Training Examination scores associated with weekly assigned reading and preparatory examinations. *Arch Surg*. 2003;138(11):1195–1197.
45. DeVirgilio C, Stabile BE. Weekly reading assignments and examinations result in sustained improvement in American Board of Surgery In-Training Examination (ABSITE) scores. *Am Surg*. 2005;71(10):830–832.
46. Ahmed M, Matolo N, Sloan D, et al. Didactic surgical education by faculty: the effect on American Board of Surgery In-Training examination percentile scores. *Am Surg*. 2006;72(12):1176.
47. Mahmoud A, Andrus CH, Matolo NM, Ward CC. Directed postgraduate study result in quantitative improvement in American Board of Surgery In-Training Exam scores. *Am J Surg*. 2006;191(6):812–816.
48. Borman KR. Does academic intervention impact ABS qualifying examination results? *Curr Surg*. 2006;63(6):367–372.
49. Gregg SC, Eisenberg D, Duffy AJ, Longo WE. Design, management, and critical evaluation of a surgical basic/clinical science curriculum: the role of an educational chief resident. *J Surg Educ*. 2008;65(1):36–42.
50. Farrohki ET, Jensen AR, Brock DM, et al. Expanding resident conferences while tailoring them to level of training: a longitudinal study. *J Surg Educ*. 2008;65(2):84–90.
51. Kosir MA, Fuller L, Tyburski J, et al. The Kolb learning cycle in American Board of Surgery in-training exam remediation: the accelerated clinical education in surgery course. *T Am J surg*. 2008;196(5):657–662.
52. Lube MW, Borman KR, Fulbright AE, Friedell ML. Retrospective evaluation of residents' American Board of Surgery In-Service Training Examination (ABSITE) scores as a tool to evaluate changes made in a basic science curriculum. *J Surg Educ*. 2010;67(3):167–172.
53. Corneille MG, Willis R, Stewart RM, Dent DL. Performance on brief practice examination identifies residents at risk for poor ABSITE and ABS qualifying examination performance. *J Surg Educ*. 2011;68(3):246–249.
54. Willis RE, Peterson RM, Dent DL. Usefulness of the American College of Surgeons' Fundamentals of Surgery Curriculum as a knowledge preparatory tool for incoming surgery interns. *Am J Surg*. 2013;205(2):131–136.
55. Krajewski A, Filippa D, Staff I, et al. Implementation of an intern boot camp curriculum to address clinical competencies under the new Accreditation Council for Graduate Medical Education supervision requirements and duty hour restrictions. *JAMA Surg*. 2013;148(8):727–732.
56. Blackmore C, Austin J, Lopushinsky SR, Donnon T. Effects of postgraduate medical education "boot camps" on clinical skills, knowledge, and confidence: a meta-analysis. *J Grad Med Educ*. 2014;6(4):643–652.
57. Buckley EJ, Markwell S, Farr D, et al. Improving resident performance on standardized assessments of medical knowledge: a retrospective analysis of interventions correlated to American Board of Surgery In-Service Training Examination performance. *Am J Surg*. 2015;210(4):734–738.
58. Kelly DM, London DA, Siperstein A, et al. A structured educational curriculum including online training positively impacts American Board of Surgery In-Training Examination scores. *J Surg Educ*. 2015;72(5):811–817.
59. Willis RE, Dent DL, Love JD, et al. Predicting and enhancing American Board of Surgery In-Training Examination performance: does writing questions really help? *Am J Surg*. 2016;211(2):361–368.
60. Lamb LC, DiFiori MM, Jayaraman V, et al. Gamified twitter microblogging to support resident preparation for the american board of surgery In-Service training examination. *J Surg Educ*. 2017;74(6):986–991.
61. Tarabichi S, DeLeon M, Krumrei N, et al. Competition as a means for improving academic scores and attendance at education conference. *J Surg Educ*. 2018;75(6):1437–1440.
62. Kantar RS, Wise E, Morales D, et al. The American board style practice in-training examinationasa predictor of performance on the AmericanBoard of surgery in-training examination. *J Surg Educ*. 2018;75:895–900.