Academic outcomes of flipped classroom learning: a meta-analysis

Kuo-Su Chen,[1,2,3] Lynn Monrouxe,[2,3] Yi-Hsuan Lu,[1,2,3] Chang-Chyi Jenq,[2,3,4] Yeu-Jhy Chang,[2,3,5] Yu-Che Chang[2,3,6] & Pony Yee-Chee Chai[7]

CONTEXT The flipped classroom (FC), reversing lecture and homework elements of a course, is popular in medical education. The FC uses technology-enhanced pre-class learning to transmit knowledge, incorporating in-class interaction to enhance higher cognitive learning. However, the FC model is expensive and research on its effectiveness remains inconclusive. The aim of this study was to compare the efficacy of the FC model over traditional lecture-based (LB) learning by meta-analysis.

METHODS We systematically searched MEDLINE, PubMed, ERIC, CINAHL, EMBASE, reference lists and Association for Medical Education in Europe (AMEE) conference books. Controlled trials comparing academic outcomes between the FC and LB approaches in higher education were considered eligible. The main findings were pooled using a random-effects model when appropriate.

RESULTS Forty-six studies (9026 participants) were included, comprising four randomised controlled trials (RCTs), 19 quasi-experimental studies and 23 cohort studies. Study populations were health science (n = 32) and non health science (n = 14) students. The risk of bias was high (36/37 articles). Meta-analyses revealed that the FC had significantly better outcomes than the LB method in examination scores (post-intervention and pre–post change) and course grades, but not in objective structured clinical examination scores. Subgroup analyses showed the advantage of the FC was not observed in RCTs, non-USA countries, nursing and other health science disciplines and earlier publication years (2013 and 2014). Cumulative analysis and meta-regression suggested a tendency for progressively better outcomes by year. Outcome assessments rarely focused on behaviour change.

CONCLUSIONS The FC method is associated with greater academic achievement than the LB approach for higher-level learning outcomes, which has become more obvious in recent years. However, results should be interpreted with caution because of the high methodological diversity, statistical heterogeneity and risk of bias in the studies used. Future studies should have high methodological rigour, a standardised FC format and utilise assessment tools evaluating higher cognitive learning and behaviour change to further examine differences between FC and LB learning.

Medical Education 2018: 52: 910–924
doi: 10.1111/medu.13616

[1]Department of Nephrology, Chang Gung Memorial Hospital, Keelung branch, Keelung,Taiwan
[2]Chang Gung Medical Education Research Center (CG-MERC), Chang Gung Memorial Hospital, Linkou Branch, Taoyuan,Taiwan
[3]College of Medicine, Chang Gung University, Taoyuan,Taiwan
[4]Department of Nephrology, Chang Gung Memorial Hospital, Linkou branch, Taoyuan, Taiwan
[5]Department of Neurology, Chang Gung Memorial Hospital, Linkou branch, Taoyuan, Taiwan
[6]Department of Emergency Medicine, Chang Gung Memorial Hospital, Linkou branch, Taoyuan, Taiwan
[7]Department of Pharmacy, Chang Gung Memorial Hospital, Keelung branch, Keelung, Taiwan

Correspondence: Lynn Valerie Monrouxe, Chang Gung Medical Education Research Centre (CG-MERC), Chang Gung Memorial Hospital, Taoyuan, Taiwan Tel: 00 886 3 328 1200; E-mail: monrouxe@me.com

INTRODUCTION

The flipped classroom (FC) is a hybrid approach, combining online learning and face-to-face classroom activities. In this pedagogical model, students engage in content learning *before* class, thereby maximising in-class time for active learning.[1] The FC utilises technology for pre-class learning, with face-to-face classrooms becoming interactive learning activities. This methodology restructures and reorders traditional lecture-based (LB) approaches by moving students, rather than teachers, to the centre of learning.[2–4] Such active learning should improve outcomes as learners practise with, engage with and apply their pre-class learning.[5] Although modern versions of the FC appeared over 10 years ago,[6] and despite its popularity in education generally and medical education specifically, we lack firm conclusions regarding its efficacy.[7,8]

The FC approach is underpinned by active learning theory.[9] Active learning has been defined as 'any instructional method that engages students in the learning process'.[10] An active learning approach is one in which students undertake meaningful learning activities, cognitively engaging in those activities. Previous research suggests that students' understanding and performance are improved via active learning.[10–13] A meta-analysis confirmed that active learning increased student performance and decreased failure rate in undergraduates.[14] Thus, as FC is a form of active learning,[4,15] it has been associated with better learning outcomes,[16] especially in higher-order thinking.[17,18] Indeed, research suggests that an FC approach increases motivation, satisfaction, performance and class attendance rate.[4,19,20] However, research results are generally inconsistent,[2,19,21–29] even when the same outcome measure is considered.[7,8]

A number of systematic reviews have been undertaken in the area of health care professionals' education that highlight this issue of inconsistency. For example, in a systematic review of the effectiveness of the FC in medical education, Chen et al.[7] found nine articles reporting randomised or non-randomised controlled studies and a further 37 reporting studies with 'other' methods or descriptive accounts (e.g. action research, pre–post designs and commentaries). Although students' perceptions of the FC process were typically positive across all nine of the controlled studies, Chen et al.,[7] reported mixed results in terms of changes in knowledge, and skills. For example, regarding outcomes using multiple-choice questions (MCQs), some studies reported positive findings for the FC over the LB approach,[30,31] whereas others found no difference.[22,32] Furthermore, Chen et al. found a varying direction and magnitude of the effect sizes and confidence intervals (CIs) across studies.[7] A more recent systematic review by Hughes and Lyons[8] also found mixed results across 11 studies when considering MCQ outcomes: four studies each demonstrated an advantage and a disadvantage of the FC approach over the LB approach and a further three studies showed mixed results.

Systematic reviews have also been undertaken in the area of nurse education, reporting similar ambiguous results. For example, Betihavas et al.[33] identified only five studies for inclusion in their systematic review. They found that although the FC condition tended to produce better outcomes immediately post-event[26,34] alongside positive course evaluations,[35,36] when it came to performance in the final examination, any advantage was lost.[26] Furthermore, not all studies reported an advantage for the FC,[2] and not all course evaluations were positive.[34] Betihavas et al.[33] concluded that there was a lack of evidence for the efficacy of the FC, claiming that much could be learned by examining how 'other health disciplines, such as pharmacy, have contributed extensively to implementing flipped classroom models'. Furthermore, in a narrative review of 13 articles reporting on the FC method in nurse education, Presti[20] concluded that 'few studies exist on the flipped classroom in nursing education, and only one study statistically validates its value in improving examination scores'. Finally, a recent integrative review also found 13 empirical studies reporting FC outcomes in nursing education (with some overlap in studies reported by Betihavas et al[33] and Presti[20]).[27] Unsurprisingly, it too concluded that results from FC studies were mixed.

As we can see from these literature reviews, we are left with a rather muddy picture regarding the efficacy of FC over LB learning. Given the importance of the FC, it seems pertinent that we employ research methods to further understand whether the FC method works and in what circumstances. Data aggregation by meta-analysis[37–39] can settle controversies from conflicting studies, allowing greater power and precision to estimate the effectiveness of relative interventions, and can generate new hypotheses. In a meta-analysis, the

cause of inconsistency between studies can be explored through subgroup analyses and meta-regression. Thus, a meta-analysis may help to answer questions that individual studies cannot.

Following this notion, two meta-analyses have very recently been published by researchers from China, drawing on studies examining the FC method in nursing.[40,41] Hu et al.[40] focused purely on randomised controlled trials (RCTs) of the FC method with baccalaureate nursing students in China between 2015 and 2017. They found 11 RCTs matching their inclusion and exclusion criteria, which measured two broad outcomes: theoretical knowledge, and skills. In terms of theoretical knowledge, only one of the nine studies reporting this outcome demonstrated no difference between the FC and LB approaches. Results of meta-analyses showed a significant difference between the FC and lecture, favouring the FC approach. For the outcome of students' skills scores, only five studies were included in the meta-analysis, with only one study demonstrating no difference between learning conditions. Again, the pooled effect size found an overall significant difference favouring the FC over the LB condition. Tan et al.[41] also undertook a meta-analysis of studies reporting the relative effectiveness of the FC in Chinese nursing education (2014–2016). They included 29 articles in their review, measuring theoretical examination scores (n = 16), skills examination scores (n = 16) and measures of self-learning via questionnaires (n = 15). Briefly, they found significant differences in favour of the FC approach for all three outcome measures.[41]

Despite the number of recent reviews examining the relative efficacy of the FC over the LB approach, a gap in the literature remains: all of the reviews were limited to a single discipline (e.g. nursing education[33,40,41] and medical education[7,8]), with a small number of articles generally favouring the FC approach (often between five and 11, with only one including 29), mostly being narrative syntheses with only two meta-analyses.[40,41] These two meta-analyses considered data from a single country (China), only pooling RCTs from Chinese nursing education published in Chinese journals. The results of such meta-analyses from this very restricted data source may be difficult to generalise.

Given that the FC pedagogical model has been adopted across a wide range of disciplines, data pooling across multiple disciplines may help to increase statistical power and identify differences in the effects across different domains. Thus, a more inclusive approach, not restricted to a single discipline or single country, can open up the possibility of multiple subgroup analyses enabling an understanding of the nuances around the relative efficacy of the FC approach in comparison with LB learning. We therefore developed a study using a meta-analysis technique that was not limited to a single discipline or country, to answer the following research questions (RQs).

RQ1 Is the FC learning approach associated with students' enhanced knowledge and behaviour more than the LB learning approach in higher education?

RQ2 Does this differ by context (e.g. research method, publication year, publication forum, student group and learning environment)?

METHODS

The meta-analysis was undertaken following the PRISMA (*p*referred *r*eporting *i*tems for *s*ystematic reviews and *m*eta-*a*nalyses) statement (see Box 1 for definitions of terms used in a meta-analysis).[42]

Eligibility criteria

The criteria for study eligibility are listed in Table 1 and include: one comparator study examining the FC against the LB approach; undergraduate students or higher; assessment of learning, behaviour change and impact of behaviour change, rather than student satisfaction,[43] and outcomes quantitatively measured. Duplicate reports were excluded.

Data sources

The search took place in June 2016. Data sources included electronic databases, reference lists, conference abstracts and dissertations. Electronic databases included MEDLINE, PubMed, the Cumulative Index to Nursing and Allied Health Literature (CINAHL), the Education Resources Information Center (ERIC) and Excerpta Medica Database (EMBASE). Related publications that appeared on the web page during the electronic search (e.g. 'similar articles' appearing on the screen in PubMed, or 'other users also viewed this article' appearing in ScienceDirect) were enrolled if eligible. Reference lists of relevant manuscripts were searched for additional literature. Theses or

Box 1 Definition of terms in a meta-analysis

Meta-analysis
A statistical analysis that combines the results of multiple quantitative information from related studies and produces results that summarise a whole body of research.

Begg's test
An indicator of publication bias. Begg et al. proposed testing the interdependence of variance and effect size using Kendall's method. This bias indicator makes fewer assumptions than that of Egger et al. but it is insensitive to many types of bias to which Egger's test is sensitive.

Cochrane *Q*
A classical measure of heterogeneity in meta-analysis. *Q* is distributed as a chi-squared statistic with $n - 1$ degrees of freedom. For the *Q*-statistic test, a 'Cochrane *Q*-value > degree of freedom' or a 'p-value of < 0.10' suggests the presence of heterogeneity.

Egger's test
Another indicator of publication bias. Egger et al. suggested a test to examine asymmetry in funnel plots. This is a test for the Y-intercept = 0 from a linear regression of the normalised effect estimate (i.e., the estimate divided by its standard error) against precision (the reciprocal of the standard error of that estimate).

Funnel plot
A graph that provides a visual check of the existence of publication bias. In the absence of publication bias, it assumes that studies with high precision will be plotted near the average, and studies with low precision will be spread evenly on both sides of the average, creating a roughly funnel-shaped distribution. Asymmetry in funnel plots may indicate the presence of publication bias in meta-analysis.

Heterogeneity
Heterogeneity in meta-analysis refers to the variation or inconsistency in the study outcomes between studies.

***I*-squared (I^2)**
Another measure of heterogeneity in meta-analysis. The I^2 statistic is the percentage of variation across studies attributable to heterogeneity. The formula of $I^2 = 100\% \times (Q\text{–df})/Q$. An I^2 of < 25% represents low heterogeneity, 25–50% moderate heterogeneity and > 50% high heterogeneity.

Meta-regression
A statistical method that can be used in meta-analysis to examine the impact of moderator variables on study effect size using regression-based techniques.

PRISMA statement
A guideline for the reporting of systematic reviews and meta-analyses. The PRISMA statement comprises 27 items and a four-phase flow diagram.

Publication bias
Occurs when the published results depend not just on the quality of the research but also on the hypothesis tested, and the significance and direction of effects detected. It suggests that results not supporting the hypotheses of researchers often go unpublished, leading to a bias in published research.

Random-effects model
The process of meta-analysis is undertaken by either a fixed-effect or a random-effects statistical model. A fixed-effect meta-analysis assumes all studies are estimating the same (fixed) treatment effect, whereas a random-effects meta-analysis permits differences in the treatment effect across studies.
In a random-effects model, the observed heterogeneity is attributed to two sources: (i) between-study heterogeneity in true effects, and (ii) within-study sampling error.

Table 1 Inclusion and exclusion criteria

Inclusion criteria	Exclusion criteria
Two groups, controlled studies (randomised or non-randomised, interventional or observational)	Reviews, editorials, qualitative studies or single-arm before–after studies
Research question meets the following PICO (*P*roblem, *I*ntervention, *C*omparison, *O*utcome) criteria:	Outcome only reporting satisfaction or perception survey
(P) higher education (any discipline, any level)	K-12 education
(I) flipped classroom learning in any format	Control group not traditional lecture-based learning
(C) traditional lecture-based learning	Duplicate reporting (studies from same author or institution)
(O) academic performance (any Kirkpatrick level 2 to level 4 learning outcome)[27]	Self-rated or self-assessed academic outcomes
Quantitative outcomes data available	Pre-intervention assessment or assessment during the process of intervention, such as quiz or homework
Post-intervention assessment, change in pre–post intervention assessment	
Any language	

conference abstracts appearing in the reference lists or the electronic searches were included if eligible. The Association for Medical Education in Europe (AMEE) conference books (2007–2016) were searched for unpublished studies.

Search strategy

The key search terms used in the electronic database searches were 'flipped classroom', 'flipped education', 'flipped learning', 'reverse classroom', 'backward classroom', 'inverted classroom' and 'inverse classroom'.

Study selection

Study selection was based on the a priori set of criteria for inclusion and exclusion: for articles that clearly met the inclusion criteria, or 'possibly' met them, full-text publications were retrieved for the formal review. Each study was independently assessed by at least two authors (Y-CC, C-CJ and Y-JC). All disagreements between them were resolved by discussion or third-party adjudication (K-SC).

Assessment of methodological quality

The Effective Public Health Practice Project (EPHPP) Quality Assessment Tool[44] was used to appraise methodological quality as it is suitable for both interventional (RCT, quasi-experimental) and observational (cohort, historical control and case–control) studies. With this tool, each study was rated as strong, moderate or weak according to the following components: selection bias; study design; confounding factors; study blinding; data collection; withdrawals; and dropouts. A global rating was then allocated using the standard system: strong (no weak ratings); moderate (one weak rating); or weak (two or more weak ratings). Risk of bias was only assessed in journal articles and theses. Conference abstracts were not appraised for risk of bias because of the limited information available. Risk of bias was assessed by two authors independently (Y-HL and PY-CC). Disagreements between them were again resolved by discussion or third-party adjudication (K-SC).

Data synthesis

Data were independently extracted by two reviewers (Y-HL and K-SC) using a standardised form. The extracted information included publication type and year, design, demographics of participants, intervention and control, and type of outcomes measured. The cognitive levels of learning outcomes measured were categorised according to the Kirkpatrick model[43] and Bloom's taxonomy.[45]

Given that multiple effect measures are usually reported in educational research,[46–49] several methods have been proposed to manage multiple effect measures in meta-analysis.[50] In this work, we pooled data separately based on the type of effect measure. When several similar effect measures were reported in the same study, we selected the one

that best represented the particular type of outcome required for the meta-analysis.[50] Data pooling was undertaken by combining the standardised mean difference (SMD) using a random-effects model.[51] If a study reported several datasets for the same type of effect measure, the averaged mean and pooled variance was used for data pooling.[52] For studies that did not report standard deviations (SDs) for their estimates, or had other missing data, we contacted the author by e-mail first. If no response from authors was forthcoming, we synthesised the SD from the CIs or exact p-values, interquartile ranges and minimum–maximum by the principle of estimation method as previously reported.[53–55] We eliminated studies from data pooling if the required information for meta-analysis (such as participant number, mean and SD) could not be obtained.

Heterogeneity of effect sizes across studies was evaluated using the *Q*-statistic and quantified by *I*-squared values.[56] Subgroup analyses were performed to explore the impacts of different contexts on the outcome measures. Cumulative analyses and meta-regression were undertaken to examine the relationship between publication year or sample size and effect size. Publication bias was evaluated through visual inspection of funnel plots and Egger's regression test. Meta-analyses were carried out using STATA 11.2 software (StataCorp LP, College Station, TX, USA).

RESULTS

Study search and selection

The process is summarised in a PRISMA[42] flow diagram (Fig. 1). The electronic search resulted in 1682 potentially relevant citations. An additional 24 studies were added through another source. These 1706 retrieved records were pooled into a single database in EndNote (Clarivate Analytics, Philadelphia, PA, USA). After removing 428 duplicates, 1278 citations were subject to title and abstract screening and 114 studies were subject to a full-text review. Finally, 46 studies with a total of 9026 participants were included. The characteristics of these studies are summarised in Table S1.

Scope of the included studies

Of the 46 included studies, 36 were journal articles, nine were conference abstracts and one was a thesis (Table S1). In terms of design, they included four RCTs, 19 quasi-experimental studies and 23 cohort studies. Thirty-seven studies originated from the USA and nine from other countries. There were 32 health science and 14 non-health science education studies. Among the 32 health science education studies, 14 involved medical education, nine pharmacy education, three nursing education and six were from other health science education disciplines.

Outcomes measured

Multiple effect measures were reported. We classified the effect measures into three categories: examination scores ($n = 41$); course grades ($n = 9$); and objective structured clinical examinations (OSCEs) ($n = 2$). The examination scores were the results of MCQ tests ($n = 20$) and written examinations or essays ($n = 3$), a combination of different formats ($n = 5$) or were unclear ($n = 13$). Of the 41 studies reporting examination scores, seven were before–after studies and provided pre–post score changes. All examination score outcomes evaluated knowledge change according to Kirkpatrick's level 2 measurement (learning).[43] The level of cognitive learning evaluated in these tests comprised only lower cognitive levels in one study, ranging from low to high (levels 1–6 in Bloom's taxonomy[45]) in 16 studies and were not described in 24 studies. Multiple assessments, such as homework assessments, attendance rates, quizzes, presentations, and mid-term and final examination scores, were combined for the calculation of a *grade score*. The formula for the course grade calculation differed between studies.

Quality assessment of the included studies

Methodological quality assessment (36 journal articles, one thesis) showed that 36 studies exhibited a high risk of bias and one study had a moderate risk of bias (see Table S2 for details). The weakest indicator was blinding. Risks were also high in the data collection, confounders and withdrawal domains.

Poor reporting quality was an important cause of a high risk of bias. Many quality indicators were not clearly described or were absent. Frequently missing information included: study design; eligible population; participant and dropout populations; important confounders between groups at baseline, and the reliability/validity of outcome assessment tools. This resulted in many studies being rated with a high risk of bias.

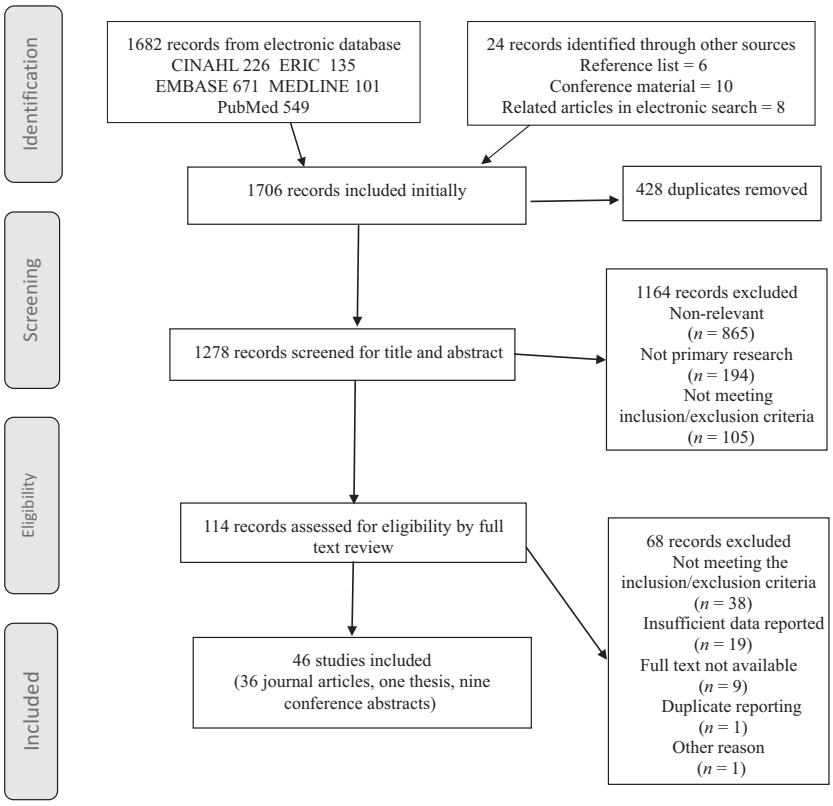

Figure 1 PRISMA flow chart of the study selection process

Main and subgroup meta-analyses

We now report the main and subgroup meta-analyses to address our two related RQs. Is the FC learning approach associated with better academic performance than the LB learning approach in higher education? Does this differ by context (e.g. research method, publication year, publication forum, student group and learning environment)? Figures S1–S9 are available to visually support our interpretation and we will refer to them throughout. Following Valentine et al.,[57] we note that it is possible to undertake a meta-analysis with as few as $n = 2$ studies.

Meta-analyses by outcome measures

Data pooling of all 46 studies was undertaken according to the type of outcome measure reported. Significantly higher examination scores, pre–post examination score improvements and course grades for the FC approach were identified (Table 2, Fig. 2). However, no significant effect was noted when the outcome measure was an OSCE score. Statistical heterogeneity was very high for all types of outcome measured.

Data pooling for the 32 health science education studies showed similar results to the main analysis: the FC condition had higher examination scores, higher course grades and better pre- to post-test changes, but similar OSCE scores (Table 2, Fig. S1). Data pooling for the 14 studies in non-health science education showed a better outcome for the FC condition in examination scores, but not in the pre- to post-test scores or course grades (Table 2, Fig. S2).

Detailed analyses regarding the outcomes reported for medical education, pharmacy education, nursing education and other disciplines in health sciences education separately by outcome measure were undertaken (see Table 2 and Figs S2–S6 for all details). Briefly, examination scores in the FC condition were significantly better in medical and pharmacy education but not in nursing and other health disciplines, and course grades were significantly better in pharmacy education. We were not able to undertake an analysis for some outcomes because of a lack of studies.

Subgroup analyses by context

A subgroup analysis was undertaken for the outcome of examination scores. This showed an

Table 2 Meta-analyses according to discipline and academic outcomes

	Test for heterogeneity			Test for effect	
Type of outcome measured	**I^2, %**	***Q*-statistic**	**p-value**	**Pooled effect size (95% CI)**	**p-value**
All studies (*n* = 46)*					
Examination scores (*n* = 41)	90.5	422.42	<0.001	0.468 (0.307, 0.629)	<0.001
Pre–post-test change (*n* = 7)	82.6	34.54	<0.001	0.454 (0.139, 0.769)	0.005
Course grade (*n* = 9)	76.0	33.33	<0.001	0.354 (0.157, 0.552)	<0.001
OSCE (*n* = 2)	98.7	74.18	<0.001	3.116 (−2.219, 8.451)	0.252
Health science studies (*n* = 32)*					
Examination scores (*n* = 28)	86.9	205.67	<0.001	0.451 (0.277, 0.625)	<0.001
Pre–post-test change (*n* = 5)	82.3	22.58	<0.001	0.603 (0.264, 0.943)	<0.001
Course grade (*n* = 5)	49.4	7.91	0.095	0.440 (0.245, 0.636)	<0.001
OSCE (*n* = 2)	98.7	74.18	<0.001	3.116 (−2.219, 8.451)	0.252
Non-health science studies (*n* = 14)*					
Examination scores (*n* = 13)	94.4	216.19	<0.001	0.490 (0.143, 0.837)	0.006
Pre–post-test change (*n* = 2)	33.2	1.50	0.221	−0.022 (−0.425, 0.381)	0.916
Course grade (*n* = 4)	81.8	16.50	0.001	0.274 (−0.080, 0.629)	0.130
Medical education studies (*n* = 14)*					
Examination scores (*n* = 14)	80.5	66.59	<0.001	0.53 (0.31, 0.74)	<0.001
Pre–post-test change (*n* = 4)	80.2	15.15	0.002	0.71 (0.31, 1.10)	<0.001
OSCE (*n* = 1)				5.86 (4.77, 6.95)	
Pharmacy education studies (*n* = 9)*					
Examination scores (*n* = 7)	92.5	80.39	<0.001	0.53 (0.12, 0.93)	0.011
Course grade (*n* = 3)	35.9	3.12	0.210	0.53 (0.35, 0.71)	<0.001
Nursing education studies (*n* = 3)*					
Examination scores (*n* = 2)	79.0	4.76	0.029	0.20 (−0.33, 0.73)	0.468
Course grade (*n* = 1)				0.05 (−0.38, 0.49)	
OSCE (*n* = 1)				0.41 (−0.17, 1.00)	
Other disciplines in health science (*n* = 6)*					
Examination scores (*n* = 5)	89.4	37.74	<0.001	0.20 (−0.34, 0.74)	0.470
Pre–post-test change (*n* = 1)				0.26 (0.03, 0.49)	
Course grade (*n* = 1)				0.29 (−0.21, 0.78)	

*Some studies in this classification had more than one outcome measure. CI = confidence interval.

advantage of the FC condition in medical, pharmacy and non-health science education studies, but not in nursing education and other disciplines in health science education (see Table 3 for all subgroup analyses results). The subgroup analysis by study design showed that the advantage of the FC condition was observed in cohort and quasi-experimental studies, but not in RCTs. Of the four RCTs, only one conference study with small participant numbers from Thailand[58] suggested an advantage for the FC approach.

For publication year, no significant differences were found in the outcome scores for 2013 and 2014, but higher examination scores for the FC condition were found in 2015 and 2016. Cumulative analysis by year (Fig. S7) revealed a progressive improvement in outcome scores for the FC condition over time. In terms of publication forum, both studies published in journals and conference studies showed a significantly better effect for the FC method.

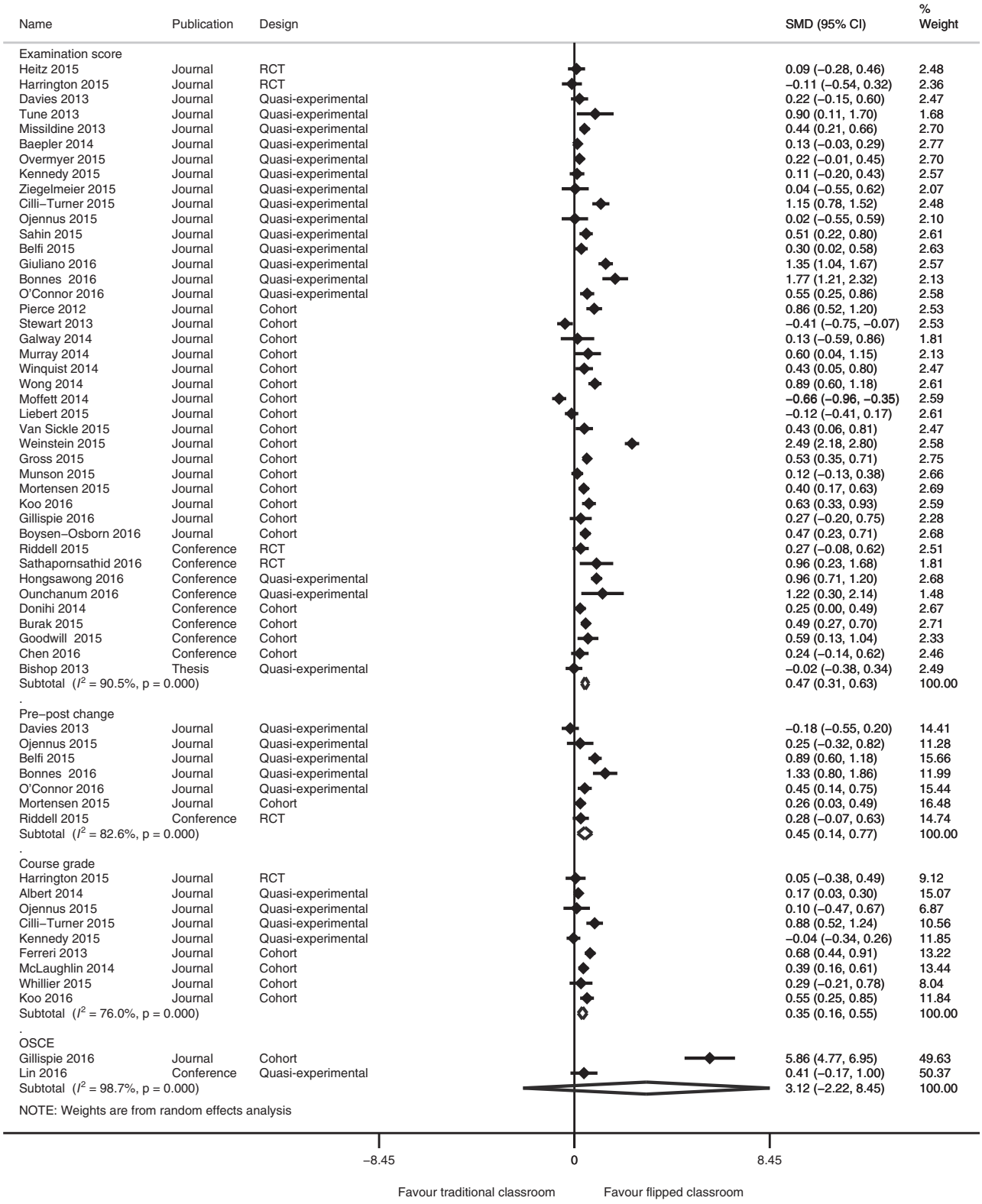

Figure 2 Meta-analyses of all 46 studies by outcome measure. CI = confidence interval; RCT = randomised controlled trial; SMD = standardised mean difference

Table 3 Subgroup analyses for post-learning examination score outcomes stratified by various contexts

	Test for heterogeneity			Test for effect	
Subgroup and domain	**I^2, %**	***Q*-statistic**	**p-value**	**Pooled effect size (95% CI)**	**p-value**
Category of discipline (*n* = 41)					
Medical (*n* = 14)	80.5	66.59	<0.001	0.527 (0.311, 0.744)	<0.001
Pharmacy (*n* = 7)	92.5	80.39	<0.001	0.527 (0.122, 0.932)	0.011
Nursing (*n* = 2)	79.0	4.76	0.029	0.196 (−0.333, 0.726)	0.468
Other health science (*n* = 5)	89.4	37.74	<0.001	0.200 (−0.343, 0.744)	0.470
Non-health science (*n* = 13)	94.4	216.19	<0.001	0.490 (0.143, 0.837)	0.006
Study design (*n* = 41)					
Cohort (*n* = 20)	93.3	284.74	<0.001	0.435 (0.173, 0.696)	0.001
Quasi-experimental (*n* = 17)	87.2	125.27	<0.001	0.548 (0.328, 0.768)	<0.001
RCT (*n* = 4)	54.6	6.61	0.085	0.219 (−0.108, 0.545)	0.190
Publication year (*n* = 41)					
2012 (*n* = 1)				0.858 (0.517,1.198)	
2013 (*n* = 5)	80.7	20.69	<0.001	0.168 (−0.205, 0.541)	0.376
2014 (*n* = 7)	89.5	56.99	<0.001	0.245 (−0.115, 0.606)	0.182
2015 (*n* = 18)	92.5	226.45	<0.001	0.426 (0.167, 0.685)	0.001
2016 (*n* = 10)	81.8	49.58	<0.001	0.801 (0.526, 1.075)	<0.001
Publication type (*n* = 41)					
Journal (*n* = 32)	92.2	388.90	<0.001	0.458 (0.263, 0.653)	<0.001
Thesis (*n* = 1)				−0.018 (−0.379,0.344)	
Conference (*n* = 8)	72.0	25.00	0.001	0.543 (0.306, 0.780)	<0.001
Country (*n* = 41)					
USA (*n* = 34)	90.5	348.56	<0.001	0.476 (0.304, 0.649)	<0.001
Non-USA (*n* = 7)	91.9	73.65	<0.001	0.435 (−0.064, 0.935)	0.088
Format of test (*n* = 41)					
MCQ (*n* = 20)	87.4	150.97	<0.001	0.394 (0.186, 0.601)	<0.001
Written essay (*n* = 3)	93.7	31.93	<0.001	1.186 (0.451, 1.921)	0.002
Combination (*n* = 5)	51.6	8.26	0.082	0.378 (0.174, 0.582)	<0.001
Unclear (*n* = 13)	93.9	195.40	<0.001	0.451 (0.087, 0.815)	0.015
Setting of medical education studies (*n* = 14)					
School (*n* = 7)	66.7	18.02	0.006	0.578 (0.342,0.814)	<0.001
Hospital (*n* = 7)	85.5	41.42	<0.001	0.489 (0.123,0.856)	0.009

CI = confidence interval; MCQ = multiple-choice question; RCT = randomised controlled trial.

An advantage of the FC over the LB condition was present in studies originating from the USA, but not in studies from other countries. The advantage of the FC approach was also observed for all types of examination scores. Finally, we analysed the 14 medical education studies by setting. The FC condition had higher outcome scores than the LB condition in both hospital and school settings.

Meta-regression for the examination score outcome measure

Meta-regression by year (Fig. S8) showed a significant correlation (p = 0.023) between publication year and effect size (by excluding the year 2012, which had only one single study), but no significant relationship between sample size and outcome (p = 0.856).

Publication bias

A visual analysis via funnel plot (Fig. S9) revealed no obvious evidence of publication bias. Begg's test ($p = 0.456$) and Egger's regression test ($p = 0.236$) further confirmed no significant publication bias.

DISCUSSION

We undertook a meta-analysis of 46 studies comparing the FC and LB approaches to learning identified across a range of health science and non-health science educational fields. To our knowledge, this is the most comprehensive meta-analytical study to date on this topic. Our overall findings build on those from several recent systematic (narrative) reviews[7,8,20,33] and meta-analyses[40,41] investigating the effectiveness of the FC in health care educational settings. Thus, the systematic narrative reviews reporting mixed findings from individual studies across medical and nursing education typically identified a relatively small number of studies and consistently called for more research and a wider reach in terms of examining the relative efficacy of the FC method across different educational groups.[7,8,20,33] Our meta-analytical approach was adopted to avoid the subjectivity of the narrative approach to systematic reviews, and to pool the data from a large number of studies to ascertain the relative effectiveness of the FC method across a number of domains.[59]

Our work also builds on and extends the findings from the two meta-analyses previously discussed.[40,41] Both of these studies found a significant effect for increased learning in Chinese nursing students using the FC method in terms of theoretical knowledge,[40,41] skills[40,41] and self-directed learning.[41] We included studies across all disciplines and countries, as long as the studies were available in the electronic databases we searched, verifying the findings from the Chinese studies for examination scores and course grades (i.e. related to *theoretical knowledge*) across a range of educational settings. However, for the OSCE outcome (related to *skills*) our analysis found no advantage for the FC condition. Despite the assertion that meta-analysis can be undertaken using as few as two studies,[57] this latter finding may be a result of the small number of studies reporting OSCE outcomes in our database and the fact that their results were greatly divergent.

Our study also revealed inconsistencies in FC efficacy between different study designs: cohort or quasi-experimental studies, but not RCTs, showed a significant advantage of the FC over the LB condition. This lack of efficacy for the FC condition found in our study directly contradicts the findings from the two earlier meta-analyses,[7,8] which demonstrated a significant advantage of the FC approach in RCTs. This discrepancy may reflect, once again, the small number of RCTs in our dataset. The possibility of bias in the results of non-RCT-based studies should also be considered.

Further differences were found in our results according to the particular student group under study. Thus, medical, pharmacy and non-health science education student groups appeared to benefit from their FC experiences, but this was not the case for nursing and other health science students. Thus, our study has an advantage over the systematic reviews in medical education undertaken by Chen et al.[7] and Hughes and Lyons,[8] who concluded that results for the FC condition were mixed across studies, by pooling the data and gaining additional statistical power to ascertain the relative advantage of the FC approach within medical education settings. Indeed, the high statistical heterogeneity found in our meta-analysis actually reflects the relative inconsistency between studies, as reported by Chen et al.[7] and Hughes and Lyons.[8] In terms of nurse education, however, our study fails to replicate the findings from the two meta-analyses of the FC method using studies from the domain of nursing education in China. We are unsure exactly why this might be, although we suspect that it could be partly to do with the relative newness of the FC approach in those fields outside China,[20,33] alongside the paucity of studies in our analyses. Indeed, only three studies, distributed in different effect estimates, were incorporated in the nursing analysis. Furthermore, when considering the relative maturity of the FC methodology in each respective discipline, it is worth noting that our cumulative analyses suggest a tendency of progressive improvement in the outcome of the FC condition over time. Thus, progressive understanding of the FC method, the utilisation of newer technology, the developing maturity of teaching skills associated with the FC method and accumulated experience in the FC teaching format may all contribute to the improvement of outcomes.[7]

The results of our meta-analyses have several limitations. For example, the studies had a high degree of statistical heterogeneity. It is likely that the statistical heterogeneity is a result of the large degree of methodological diversity. In this review methodological diversity exists and takes many forms, including, as discussed above, the implementation of the FC/LB conditions and the research design used to assess the differences. As such, although all studies addressed the implementation of an FC style of education, they varied in both the format and content of classroom flipping. For example, the experiences and expertise of teachers who designed and led the FC activity, and the types and qualities of pre-class learning material were likely to differ. The content or cognitive level of knowledge evaluated (i.e. outcomes) also varied. The FC condition is designed to promote higher cognitive levels (e.g. application). As such, lower levels of assessment (e.g. recall) may not necessarily be considered as an appropriate outcome for the FC approach. The academic outcomes amongst our included studies mainly comprised three types: examination scores; OSCEs, and course grades. Most of these reflect a Kirkpatrick level 2 (learning) outcome.[43] Additionally, the OSCE may also reflect a Kirkpatrick level 3 (behaviour) outcome and the MCQ may also include questions of a higher cognitive level than mere memorisation. As the formula for assessing course grades differed between studies (as we identified in Table S2), the cognitive levels evaluated also differed. Such diversity in study designs, outcomes and target populations contributes towards the heterogeneity in effect sizes. Given that heterogeneity was high, we must be cautious in drawing sweeping conclusions around the efficacy of the FC method.

The issue of publication bias also needs to be addressed. We incorporated conference studies and a thesis in this review to avoid missing important information and publication bias. However, the methodological qualities of conference studies could not be assessed because of the limited information available. This gives these conference studies a high risk of bias. Despite this, the information obtained can be used as a reference. Given that the subgroup analysis showed a similar effect and direction of outcome between both conference and journal studies, we consider this to be positive support for our findings.

Another limitation of our meta-analysis is that the literature search was undertaken in June 2016 and because of the lengthy nature of undertaking such a detailed process (i.e. whittling down an initial 1706 identified contenders to the 46 included in this study, closely reading hundreds of papers written primarily in a non-native language, data extraction, analyses and interpretation), it has taken us 18 months to deliver our results. As the number of publications on the effectiveness of the FC approach is rapidly growing, several new articles published after June 2016 were not included in our data pool. A new meta-analysis 1 or 2 years later might yield further informative and significant findings.

However, despite these limitations, our study has strengths. We have undertaken a meta-analysis comprising data from 46 studies with a total of 9026 participants from different health care and non-health care educational disciplines across the world. This is not only the largest study of its kind that we know of, but, because it is a meta-analysis, it provides a higher level of evidence[60,61] than do the descriptive review methods presently dominating the FC literature.[7,27,33] Further, our approach has enabled us to consider the patterns within and across a variety of factors, including study population, year of study and research method, culminating in a greater understanding of the nuances around the FC method more generally than if we had restricted our data pool to one participant group alone.

Our study leads us to make future research suggestions, focusing mainly on improving study methods, research design and reporting. Thus, a standard format of classroom flipping, with well-trained and motivated facilitators, comprises the basic requirement for a comparison between FC and LB methods for future research studies. Further, given that the purpose of the FC method is to improve higher cognitive learning and promote behavioural change, greater attention should be given to ensuring that future research study designs are sensitive enough to measure the higher cognitive outcomes expected to delineate the efficacy of the FC. Indeed, the appraisal of a new pedagogy suggests that conceptual knowledge should be evaluated using novel tools, such as clinical reasoning measures.[62] Furthermore, this paradigm shift in teaching methodology may require the development of a longitudinal assessment technique that measures clinical reasoning, high-level cognitive performances and behaviour change to better realise the impact of this pedagogical intervention.[2]

Future research also needs to address the reporting of research: the poor reporting quality of educational studies is an important issue influencing the process of risk bias assessment. In many of the studies included in our meta-analysis, a lot of necessary information was absent. For example, details of the number of eligible and enrolled participants, their demographic features, dropout rate, and the validity and reliability of assessment instruments were often missing. Age and gender are important confounders in clinical studies, and this is also the case in education research. However, many publications in our review failed to mention such demographics. Similar to our findings, Horsley et al.[63] concluded that: '. . . reports of randomised studies in health professions education frequently omit elements recommended by the CONSORT statement. Most reports were assessed as having a high or unclear risk of bias.' Thus, a structured format for reporting research in medical education studies is warranted. Without such rigour in the reporting of research studies, it will be impossible for us to implement a truly evidence-based platform on which to make future decisions for curricula development.

Contributors: K-SC and LVM contributed towards the design conceptualisation. All authors participated in the data acquisition, analysis and interpretation. K-SC and Y-HL performed the statistical analysis. K-SC and LVM developed the first draft of the manuscript and led the critical revision of the paper. All authors approved the final manuscript.
Acknowledgements: none.
Funding: none.
Conflicts of interest: none.
Ethical approval: not applicable.

REFERENCES

1 DeLozier SJ, Rhodes MG. Flipped classrooms: a review of key ideas and recommendations for practice. *Educ Psychol Rev* 2017;**29** (1):141–51.
2 Harrington SA, Vanden Bosch M, Schoofs N, Beel-Bates C, Anderson K. Quantitative outcomes for nursing students in a flipped classroom. *Nurs Educ Perspect* 2015;**36** (3):179–81.
3 Zawacki A, Knutson M, Keohane EM. A student-centered active learning approach to teaching anemias in a medical laboratory science hospital-based program. *Clin Lab Sci* 2016;**29** (2):104–5.
4 McLaughlin JE, Roth MT, Glatt DM, Gharkholonarehe N, Davidson CA, Griffin LM, Esserman DA, Mumper RJ. The flipped classroom: a course redesign to foster learning and engagement in a health professions school. *Acad Med* 2014;**89** (2):236–43.
5 Taylor AT, Olofson EL, Novak WR. Enhancing student retention of prerequisite knowledge through pre-class activities and in-class reinforcement. *Biochem Mol Biol Educ* 2017;**45** (2):97–104.
6 Bergmann J, Sams A. *Flip your Classroom: Reach every Student in Every Class Every Day*. Eugene, OR: International Society for Technology in Education 2012.
7 Chen F, Lui AM, Martinelli SM. A systematic review of the effectiveness of flipped classrooms in medical education. *Med Educ* 2017;**51** (6):585–97.
8 Hughes Y, Lyons N. Does the flipped classroom improve exam performance in medical education? A systematic review. *MedEdPublish* 2017;**6** (2):38.
9 Bishop JL, Verleger MA. The flipped classroom: a survey of the research. 120th ASEE Annual Conference and Exposition; Atlanta, GA 2013.
10 Prince M. Does active learning work? A review of the research *J Eng Educ* 2004;**93** (3):223–31.
11 Armbruster P, Patel M, Johnson E, Weiss M. Active learning and student-centered pedagogy improve student attitudes and performance in introductory biology. *CBE Life Sci Educ* 2009;**8** (3):203–13.
12 Rao SP, DiCarlo SE. Active learning of respiratory physiology improves performance on respiratory physiology examinations. *Adv Physiol Educ* 2001;**25** (1–4):127–33.
13 Deslauriers L, Schelew E, Wieman C. Improved learning in a large-enrollment physics class. *Science* 2011;**332** (6031):862–4.
14 Freeman S, Eddy SL, McDonough M, Smith MK, Okoroafor N, Jordt H, Wenderoth MP. Active learning increases student performance in science, engineering, and mathematics. *Proc Natl Acad Sci U S A* 2014;**111** (23):8410–5.
15 Smith JS. Active learning strategies in the physician assistant classroom – the critical piece to a successful flipped classroom. *J Physician Assist Educ* 2014;**25** (2):46–9.
16 Jensen JL, Kummer TA, d M Godoy PD. Improvements from a flipped classroom may simply be the fruits of active learning. *CBE Life Sci Educ* 2015;**14** (1):1–12.
17 Abeysekera L, Dawson P. Motivation and cognitive load in the flipped classroom: definition, rationale and a call for research. *High Educ Res Dev* 2015;**34** (1):1–14.
18 Bayliss AJ, Warden SJ. A hybrid model of student-centered instruction improves physical therapist student performance in cardiopulmonary practice patterns by enhancing performance in higher cognitive domains. *J Phys Ther Educ* 2011;**25** (3): 14–20.
19 Evans KH, Thompson AC, O'Brien C, Bryant M, Basaviah P, Prober C, Popat RA. An innovative blended preclinical curriculum in clinical

epidemiology and biostatistics: impact on student satisfaction and performance. *Acad Med* 2016;**91** (5):696–700.
20 Presti CR. The flipped learning approach in nursing education: a literature review. *J Nurs Educ* 2016;**55** (5):252–7.
21 van Vliet EA, Winnips JC, Brouwer N. Flipped-class pedagogy enhances student metacognition and collaborative-learning strategies in higher education but effect does not persist. *CBE Life Sci Educ* 2015;**14** (3):1–10.
22 Heitz C, Prusakowski M, Willis G, Franck C. Does the concept of the 'flipped classroom' extend to the emergency medicine clinical clerkship? *West J Emerg Med* 2015;**16** (6):851–5.
23 Moffett J, Mill AC. Evaluation of the flipped classroom approach in a veterinary professional skills course. *Adv Med Educ Pract* 2014;**5**:415–25.
24 Weinstein RD. Improved performance via the inverted classroom. *Chemical Eng Educ* 2015;**49** (3):141–8.
25 Stewart DW, Panus PC, Hagemeier NE. An analysis of student performance with podcasting and active learning in a pharmacotherapy module. *Curr Pharm Teach Learn* 2013;**5** (6):574–9.
26 Geist MJ, Larimore D, Rawiszer H, Al Sager AW. Flipped versus traditional instruction and achievement in a baccalaureate nursing pharmacology course. *Nurs Educ Perspect* 2015;**36** (2):114–5.
27 Njie-Carr VP, Ludeman E, Lee MC, Dordunoo D, Trocky NM, Jenkins LS. An integrative review of flipped classroom teaching models in nursing education. *J Prof Nurs* 2017;**33** (2):133–44.
28 Johnson L, Renner J. *Effect of the Flipped Classroom Model on Secondary Computer applications Course: Student and Teacher Perceptions, Questions and Student Achievement*. Louisville, KY: University of Louisville 2012.
29 Bonnes SL, Ratelle JT, Halvorsen AJ, Carter KJ, Hafdahl LT, Wang AT, Mandrekar JN, Oxentenko AS, Beckman TJ, Wittich CM. Flipping the quality improvement classroom in residency education. *Acad Med* 2017;**92** (1):101–7.
30 Gillispie V. Using the flipped classroom to bridge the gap to generation Y. *Ochsner J* 2016;**16** (1):32–6.
31 Tune JD, Sturek M, Basile DP. Flipped classroom model improves graduate student performance in cardiovascular, respiratory, and renal physiology. *Adv Physiol Educ* 2013;**37** (4):316–20.
32 Boysen-Osborn M, Anderson CL, Navarro R, Yanuck J, Strom S, McCoy CE, Youm J, Ypma-Wong MF, Langdorf MI. Flipping the advanced cardiac life support classroom with team-based learning: comparison of cognitive testing performance for medical students at the University of California, Irvine, United States. *J Educ Eval Health Prof* 2016;**13**:11.
33 Betihavas V, Bridgman H, Kornhaber R, Cross M. The evidence for 'flipping out': a systematic review of the flipped classroom in nursing education. *Nurse Educ Today* 2016;**38**:15–21.
34 Missildine K, Fountain R, Summers L, Gosselin K. Flipping the classroom to improve student performance and satisfaction. *J Nurs Educ* 2013;**52** (10):597–9.
35 Critz CM, Knight D. Using the flipped classroom in graduate nursing education. *Nurse Educ* 2013;**38** (5):210–3.
36 Simpson V, Richards E. Flipping the classroom to teach population health: increasing the relevance. *Nurse Educ Pract* 2015;**15** (3):162–7.
37 Crowther MA, Cook DJ. Trials and tribulations of systematic reviews and meta-analyses. *Hematology Am Soc Hematol Educ Program* 2007;**1**:493–7.
38 Choi SW, Lam DMH. Trials and tribulations of a meta-analyst. *Anaesthesia* 2016;**71** (2):228–31.
39 Haidich AB. Meta-analysis in medical research. *Hippokratia* 2010;**14** (Suppl 1):29–37.
40 Hu R, Gao H, Ye Y, Ni Z, Jiang N, Jiang X. Effectiveness of flipped classrooms in Chinese baccalaureate nursing education: a meta-analysis of randomized controlled trials. *Int J Nurs Stud* 2017;**79**:94–103.
41 Tan C, Yue W-G, Fu Y. Effectiveness of flipped classrooms in nursing education: systematic review and meta-analysis. *Chin Nurs Res* 2017;**4** (4):192–200.
42 Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann Intern Med* 2009;**151** (4):264–9.
43 Kirkpatrick D. *Evaluating Training Programs: The Four Levels*. San Fransisco, CA: Berrett-Khoehler 1998.
44 Thomas BH, Ciliska D, Dobbins M, Micucci S. A process for systematically reviewing the literature: providing the research evidence for public health nursing interventions. *Worldviews Evid Based Nurs* 2004;**1** (3):176–84.
45 Bloom BS. *Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain*. New York, NY: David McKay Co 1956.
46 Ahn S, Ames AJ, Myers ND. A review of meta-analyses in education: methodological strengths and weaknesses. *Rev Educ Res* 2012;**82** (4):436–76.
47 D'Agostino JV, Murphy JA. A meta-analysis of reading recovery in United States schools. *Educ Eval Policy Anal* 2004;**26** (1):23–38.
48 Edmonds MS, Vaughn S, Wexler J, Reutebuch C, Cable A, Tackett KK, Schnakenberg JW. A synthesis of reading interventions and effects on reading comprehension outcomes for older struggling readers. *Rev Educ Res* 2009;**79** (1):262–300.
49 Tran L, Sanchez T, Arellano B, Lee Swanson H. A meta-analysis of the RTI literature for children at risk for reading disabilities. *J Learn Disabil* 2011;**44** (3):283–95.

50 Scammacca N, Roberts G, Stuebing KK. Meta-analysis with complex research designs: dealing with dependence from multiple measures and multiple group comparisons. *Rev Educ Res* 2014;**84** (3):328–64.
51 DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;**7** (3):177–88.
52 Sadeghirad B. How to combine standard deviations for three groups? ResearchGate 2014. https://www.researchgate.net/post/How_to_combine_standard_deviations_for_three_groups. [Accessed 12 May 2018.]
53 Patnaik PB. The use of mean range as an estimator of variance in statistical tests. *Biometrika* 1950;**37** (1–2):78–87.
54 Hozo SP, Djulbegovic B, Hozo I. Estimating the mean and variance from the median, range, and the size of a sample. *BMC Med Res Methodol* 2005;**5** (1):13.
55 Wiebe N, Vandermeer B, Platt RW, Klassen TP, Moher D, Barrowman NJ. A systematic review identifies a lack of standardization in methods for handling missing variance data. *J Clin Epidemiol* 2006;**59** (4):342–53.
56 Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *Br Med J* 2003;**327** (7414):557–60.
57 Valentine JC, Pigott TD, Rothstein HR. How many studies do you need? A primer on statistical power for meta-analysis. *J Educ Behav Stat* 2010;**35** (2):215–47.
58 Sathapornsathid A. Flipped classroom and its effectiveness compared with traditional-style lecture in a stroke rehabilitation medicine course for medical students: randomized controlled trials. First World Summit on Competency-based Education, Association for Medical Education in Europe, 27–28 August 2016, Barcelona.
59 Cohn LD, Becker BJ. How meta-analysis increases statistical power. *Psychol Methods* 2003;**8** (3):243–53.
60 Guyatt G, Gutterman D, Baumann MH, Addrizzo-Harris D, Hylek EM, Phillips B, Raskob G, Lewis SZ, Schünemann H. Grading strength of recommendations and quality of evidence in clinical guidelines: report from an american college of chest physicians task force. *Chest* 2006;**129** (1):174–81.
61 Hadorn DC, Baker D, Hodges JS, Hicks N. Rating the quality of evidence for clinical practice guidelines. *J Clin Epidemiol* 1996;**49** (7):749–54.
62 Benner P, Hughes RG, Sutphen M. Clinical reasoning, decisionmaking, and action: thinking critically and clinically. In: Hughes RG, ed. *Patient Safety and Quality: An Evidence-Based Handbook for Nurses*. Rockville, MD: Agency for Healthcare Research and Quality 2008.
63 Horsley T, Galipeau J, Petkovic J, Zeiter J, Hamstra SJ, Cook DA. Reporting quality and risk of bias in randomised trials in health professions education. *Med Educ* 2017;**51** (1):61–71.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article:

Figure S1. Meta-analyses of the 32 health science studies by outcome measure.
Figure S2. Meta-analyses of the 14 non-health science studies by outcome measure.
Figure S3. Meta-analyses of the medical education studies by outcome measure.
Figure S4. Meta-analyses of the pharmacy education studies by outcome measure.
Figure S5. Meta-analyses of the nursing education studies by outcome measure.
Figure S6. Meta-analyses of studies from other health science disciplines by outcome measure.
Figure S7. Cumulative analysis of examination score outcome by publication year.
Figure S8. Meta-regression between examination scores and publication year.
Figure S9. Funnel plot illustrating publication bias for the examination score outcome.
Table S1. Summary of the included studies.
Table S2. Methodological qualities of journal studies by the Effective Public Health Practice Project (EPHPP) Assessment Tool.

Received 29 November 2017; editorial comments to authors 15 March 2018; accepted for publication 10 April 2018