An Application of the Checklist for Health Economic Quality Evaluations in a Systematic Review Setting

Published in:
Value in Health

Publication status and date:
Published: 01/02/2025

DOI (link to publisher):
10.1016/j.jval.2024.10.3853

Document Version
Publisher's PDF, also known as Version of record

Document License/Available under:
Article 25fa Dutch Copyright Act

Citation for the published version (APA):
Dijk, S. W., Essafi, S., & Hunink, M. G. M. (2025). An Application of the Checklist for Health Economic Quality Evaluations in a Systematic Review Setting. *Value in Health*, *28*(2), 250-259. https://doi.org/10.1016/j.jval.2024.10.3853

ELSEVIER

ScienceDirect

Contents lists available at **sciencedirect.com**
Journal homepage: **www.elsevier.com/locate/jval**

Methodology

An Application of the Checklist for Health Economic Quality Evaluations in a Systematic Review Setting

Stijntje W. Dijk, MD, MSc, Skander Essafi, MD, Myriam G.M. Hunink, MD, PhD

ABSTRACT

Objective: Quality assessment tools serve an important role in evaluating economic evaluations. This article showcases the first application of the Checklist for Health Economic Quality Evaluations (CHEQUE) tool in a systematic review setting and offers descriptive reflections on its use.

Methods: We applied CHEQUE to 21 diverse economic evaluations in a systematic review of medical education. We visualized weighted CHEQUE scores, calculate correlations between methods and reporting sections, and provided all data and R code for reuse in future applications. Finally, we provided a detailed overview of our judgments and alternative considerations of the checklist items, and suggestions for further development.

Results: Scores ranged from 18% to 94% depending on the applied weighting method, with a positive correlation between the method and reporting quality. CHEQUE enables systematic and standardized assessment but may benefit from refinement in scoring clarity and burden reduction.

Conclusion: Our study provides insights for future guidance on applying and developing CHEQUE or similar tools in economic evaluation quality assessment.

Keywords: economics, evidence-based medicine, healthcare economics and organizations, healthcare evaluation mechanisms, medical education, medical faculty, medical students, organization and administration, systematic review as topic.

VALUE HEALTH. 2025; 28(2):250–259

Highlights

- Our study provides the first application of the Checklist for Health Economic Quality Evaluations (CHEQUE) tool in a systematic review setting.
- We demonstrated CHEQUE's utility, challenged our interpretation and decisions in assigning quality scores, and provided suggestions for the development of further guidance to use this tool.
- Our reflections on using CHEQUE provided a foundation for refining quality assessment tools to further enhance their application in systematic reviews and beyond.

Introduction

Systematic reviews play an important role in synthesizing evidence and informing decision-making processes in health economic research. Systematic reviews of economic evaluations are particularly useful as they allow readers to assess whether interventions have been demonstrated to be cost-effective, and also assess the uncertainty in the evidence base and key limitations or gaps in the evidence base.[1] Yet, a lack of standardization in systematic reviews can lead to large variations in the quality and use of economic data.[2] For readers, it is important that they can read and compare the quality of included economic evaluations. Quality assessment tools contribute to this process by providing a structured framework to appraise various aspects of study design, conduct, and reporting.

Several tools have been developed to evaluate the quality of economic evaluations,[3-18] including the well-known Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement and the *British Medical Journal* (BMJ) checklist. These tools assess various aspects of study design, conduct, and reporting, aiming to enhance transparency and rigor in economic evaluations. However, they vary in their scope, criteria, and scoring methodologies. For instance, the CHEERS statement primarily serves as a reporting tool, while the BMJ checklist and Quality of Health Economic Studies instrument delve into methodological considerations and may incorporate weighted scoring systems. Checklists have commonalities; for instance, nearly all require a clear statement on the study perspective as well as regarding sensitivity and uncertainties, and only a few address items such as ethics and distribution of ethics or the independence of investigators.[2] These differences highlight the ongoing evolution of quality assessment tools and the need for careful selection based on the specific requirements of a systematic review.

The latest quality assessment tool, the Checklist for Health Economic Quality Evaluations (CHEQUE),[19] was developed to comprehensively assess the study quality of economic evaluations. Originally piloted on a set of cost-effectiveness analyses, the tool provides a standardized approach to evaluate the methodological and reporting quality of individual studies.

In the context of improving decision sciences, the acceptance of any new tool relies on a demonstration of its practical impact through full real-world applications beyond theoretical discussions.[20] This article, therefore, provides an application of CHEQUE in a systematic review of economic evaluations in medical education. We aimed to provide insights into the challenges and opportunities that we encountered in applying the CHEQUE tool to a systematic review setting alongside ideas for future development of this guidance tool. We also provided the R code we developed to calculate weighted scores based on varied assumptions and generated summary figures.

Methods

Systematic Review Used for This Paper

The systematic review of economic evaluations in medical education was conducted between September and December 2023. The primary objective of this systematic review was to explore and present the landscape and methodology of decision analytic models and economic evaluations within the field of undergraduate medical education. It focused specifically on interventions that sought to improve medical curriculum content, methodologies, student/faculty experiences, or other aspects of medical education. The review analyzed the trade-offs between incremental benefits (eg, improved student performance and patient quality of life) and incremental costs (eg, monetary expenses, faculty time) in medical education interventions.

Studies were eligible if they concerned aspects or stakeholders of medical education, contained an intervention and comparator, and reported both an effect and cost measure. Any type of economic evaluation was eligible, including a statement of costs and effects alongside trial cost-effectiveness analyses and decision analytic modeling studies. We searched Medline, EMBASE, Web of Science, Cochrane, ERIC, Google Scholar, and Center for the Evaluation of Value and Risk Cost-Effectiveness Analyses (CEA) databases for eligible studies. Of the 6 559 identified studies, 21 met the inclusion criteria.[21-41]

The selected articles incorporated various study designs, encompassing randomized controlled trials (RCTs) (n = 12), non-RCTs (n = 4), and cross-sectional comparisons, case studies, or cohort studies (n = 4). Notably, only 1 decision analytic model was identified within the reviewed literature. The economic evaluation methodologies used included CEAs, cost-minimization studies, a case study combined with net benefit regression analysis, and a cost-benefit analysis. Some studies, however, did not explicitly explore cost-effectiveness but rather provided information on costs and effects as 2 separate outcomes.

The full details of the systematic review can be found in the International Prospective Register of Systematic Reviews (PROSPERO) database (ID: CRD42023478907).

Study Quality Assessment

We assessed the quality of studies for inclusion using the CHEQUE tool. Our assessment focused on the methodological and reporting quality of the economic evaluation within the selected articles rather than increasing the risk of bias from original trials and observational studies with inputs of economic evaluation, which would be more appropriately investigated by tools such as Risk-of-Bias (ROB) tool for randomized trials–2[42] and ROB in Nonrandomized Studies.[43] While multiple alternatives to CHEQUE exist that aim to improve the quality of economic evaluations, these were less suited for our purpose as they do not provide criteria for methodological quality assessment and/or focus on reporting.[3,11,12,14,16-18] For example, the well-known CHEERS checklist is aimed to be used as a reporting guideline, rather than an assessment of quality, and does not provide quality scores for individual articles that allow for their comparison.

All authors reviewed the CHEQUE checklist and discussed how certain scenarios and items would be judged prior to data extraction. Data extraction and quality assessment were performed collaboratively by SD and SE. Judgments were carried out non-independently to allow for case-by-case discussions while using this new checklist. Any disagreements were resolved by discussion with the research group. Final scores were decided after a final research team discussion on remaining ambiguities and uncertainties.

We generated a visual representation of the assessment using statistical software R[44] similar to the ROB visualizations from the package *robvis*.[45] The total scores were computed using CHEQUE's score weighting system. We provided both the weighted scores in which we assigned a full score for items valued as not applicable (N/A) (N/A = N/A), and excluded N/A values from both the numerator and denominator (N/A = 1). An attribute is not applicable, for example, if a question addresses a modeling aspect but the assessed study is not a decision analytic model. In addition, we calculated a percentage score by dividing the weighted score by the maximum attainable score, which, in our view, more appropriately reflects the quality of articles with many N/A items.

All articles that fulfilled the eligibility criteria were included in the review, regardless of their quality assessment score.

Correlation Plot

Following the example from Figure 2 in an article by Kim et al,[19] we compared the scores between methods (M) and reporting (R) for both N/A equals 1 and N/A equals N/A. To make this comparison, we calculated the Pearson's correlation coefficient, a statistical measure of the linear relationship between 2 variables. This coefficient ranges from -1 to $+1$ and helps quantify the strength and direction of the linear association between the scores assigned to methods and reporting. A 2-tailed P value was calculated to assess the significance of the correlation.

Results

Quality Assessment Judgements

Figure 1 illustrates the individual (Fig. 1A) and summarized scores (Fig. 1B) that were assigned to our included studies.

In Table 1,[21-41] we calculated the weighted scores for the methods and reporting sections. The mean weighted absolute score quality assessment for "methods" was 72 (range: 55-81) when assigning full scores to N/A-rated items and 36 (range: 18-70) when excluding N/A-rated items. For "reporting," these scores were 82 (range: 63-94) and 48 (range: 28-72), respectively. None of the studies obtained a full score.

The percentage quality score is the same for the N/A-full scored items (as the maximum score adds up to 100), and was 55% (range: 29%-72%) for methods, and 73% (range: 43%-90%) for reporting when excluding N/A items from the maximum obtainable score.

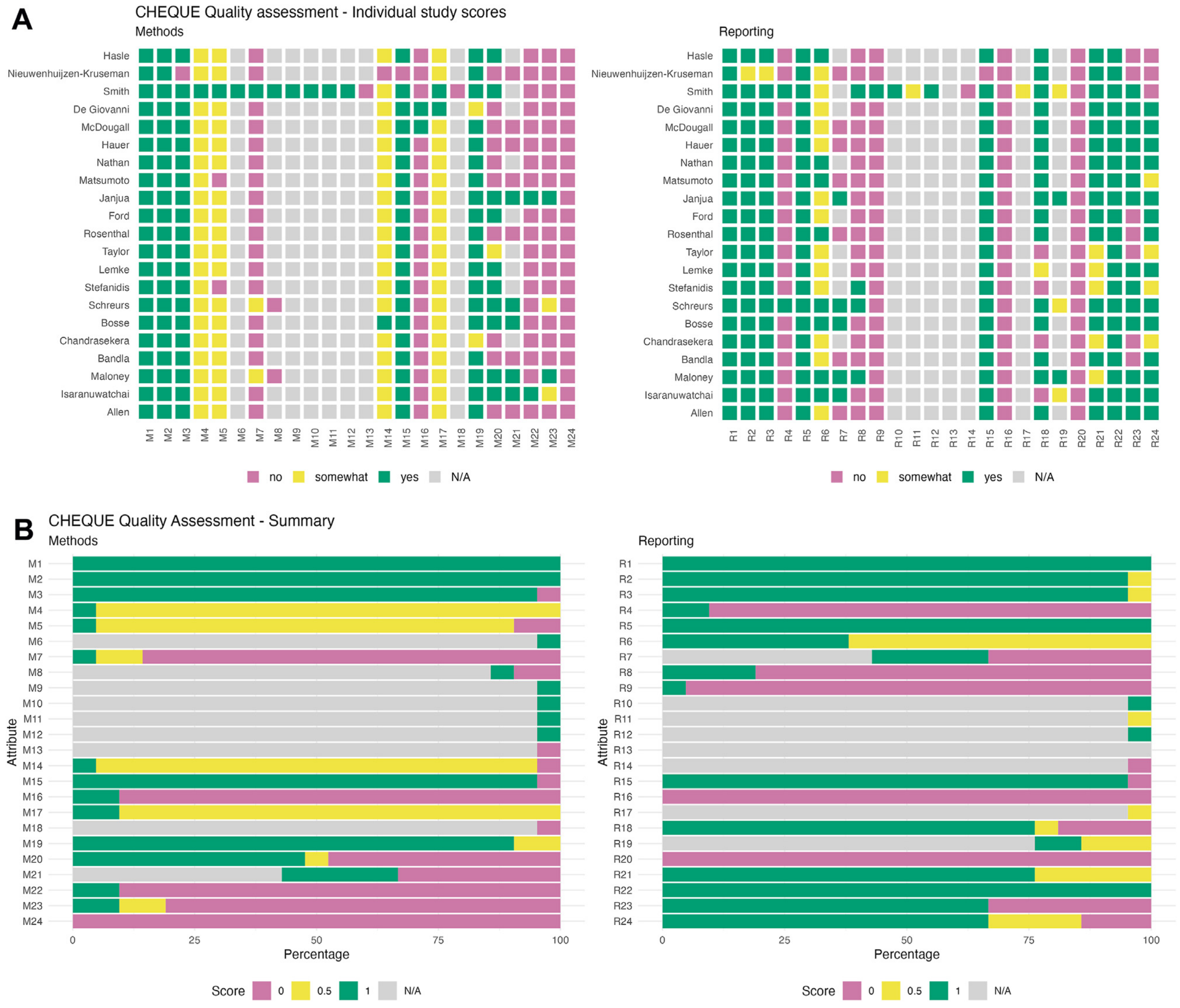
Correlations Between Methods and Reporting

We identified a positive correlation between percentage and absolute scores for methods and reporting quality. The Pearson's coefficient for the left plot in Figure 2, representing the situation in which full scores were awarded to N/A scored items, was 0.765 ($P < .001$). The middle plot, in which N/A-scored items were excluded and an absolute score was provided, had a coefficient of 0.886 ($P < .001$). Finally, the right plot, corresponding to the percentage score after the exclusion of N/A-scored items from the maximum obtainable score, had a correlation coefficient of 0.771 ($P < .001$).

Considerations, Examples, and Dilemmas

Table 2 provides an overview of the considerations we had for scoring items. We focused on examples that were outliers, partial scores ("somewhat"), or that were assigned full ("yes") or no score ("no") as we believed this may give an inaccurate impression of the study quality. For conciseness, we only highlight 1 or a limited

Figure 1. CHEQUE Quality assessment. (A) Individual study scores and (B) summary scores, separated across methods and reporting attributes.



A CHEQUE Quality assessment - Individual study scores

Methods

Reporting

B CHEQUE Quality Assessment - Summary

Methods

Reporting

N/A indicates not applicable.

number of studies per item even though the principle may be applicable to multiple studies.

Discussion

In our study, we took an initial step in applying the recently developed CHEQUE tool within the unique context of a systematic review, by providing insights and experiences that can guide future researchers. The tool, designed to assess the quality of economic evaluations, demonstrated its utility as a systematic and comprehensive instrument, encompassing both reporting and methodological aspects. Our results suggest a consistent alignment between the assessments of methods and reporting quality, signifying that higher scores in 1 domain tend to correspond with elevated scores in the other. This underscores the interconnectedness of evaluation criteria for methods and reporting, which was also found in the study by Kim et al[19] (0.837, $P = .003$) who calculated this correlation based on absolute scores.

The development process of CHEQUE involved a best-worst scaling approach among a large number of stakeholders, resulting in a well-rounded tool that assigns varying weights to different items, reflecting their contribution to the overall score. The tool offers the user flexibility, including the option to include or exclude N/A scores. However, this flexibility, though offering the user choices on the basis of what they believe is most appropriate, may also lead to differential applications and discussions on the best decision. Whereas the original CHEQUE article discusses the flexibility of handling NA scores, it does not specifically address the potential bias this introduces when comparing alongside trial CEAs to decision-analytic models. Our percentage-based scoring system offers a solution to this issue by

Table 1. CHEQUE weighted scores for methods (left) and reporting (right) items.

Study	Methods						Study	Reporting					
	N/A = full score			N/A = N/A				N/A = full score			N/A = N/A		
	Absolute score	Max score	% of max score	Score	Max score	% of max score		Absolute score	Max score	% of max score	Absolute score	Max score	% of max score
Allen et al[41]	67	100	67	30	63	47.62	Allen et al[41]	81.5	100	81.5	46.5	65	71.54
Isaranuwatchai et al[21]	79	100	79	42	63	66.67	Isaranuwatchai et al[21]	84.5	100	84.5	52.5	68	77.21
Maloney et al[22]	78	100	78	43	65	66.15	Maloney et al[22]	89.5	100	89.5	57.5	68	84.56
Bandla et al[23]	67	100	67	30	63	47.62	Bandla et al[23]	78.5	100	78.5	43.5	65	66.92
Chandrasekera et al[24]	69	100	69	29	60	48.33	Chandrasekera et al[24]	77	100	77	38	61	62.3
Bosse et al[25]	77	100	77	40	63	63.49	Bosse et al[25]	88	100	88	53	65	81.54
Schreurs et al[26]	76	100	76	41	65	63.08	Schreurs et al[26]	93.5	100	93.5	61.5	68	90.44
Stefanidis et al[27]	73.5	100	73.5	33.5	60	55.83	Stefanidis et al[27]	83	100	83	44	61	72.13
Lemke et al[28]	74	100	74	34	60	56.67	Lemke et al[28]	83	100	83	44	61	72.13
Taylor et al[29]	72	100	72	32	60	53.33	Taylor et al[29]	77	100	77	38	61	62.3
Rosenthal et al[30]	67	100	67	30	63	47.62	Rosenthal et al[30]	81	100	81	46	65	70.77
Ford et al[31]	74	100	74	34	60	56.67	Ford et al[31]	82.5	100	82.5	43.5	61	71.31
Janjua et al[32]	81	100	81	44	63	69.84	Janjua et al[32]	85.5	100	85.5	53.5	68	78.68
Matsumoto et al[33]	66.5	100	66.5	29.5	63	46.83	Matsumoto et al[33]	82	100	82	47	65	72.31
Nathan et al[34]	70	100	70	30	60	50	Nathan et al[34]	88	100	88	49	61	80.33
Hauer et al[35]	67	100	67	30	63	47.62	Hauer et al[35]	81.5	100	81.5	46.5	65	71.54
McDougall et al[36]	73	100	73	36	63	57.14	McDougall et al[36]	81.5	100	81.5	46.5	65	71.54
De Giovanni et al[37]	77.5	100	77.5	37.5	60	62.5	De Giovanni et al[37]	85.5	100	85.5	46.5	61	76.23
Smith et al[38]	73	100	73	70	97	72.16	Smith et al[38]	78.5	100	78.5	71.5	93	76.88
Nieuwenhuijzen-Kruseman et al[39]	55	100	55	18	63	28.57	Nieuwenhuijzen-Kruseman et al[39]	63	100	63	28	65	43.08
Hasle et al[40]	74	100	74	34	60	56.67	Hasle et al[40]	81	100	81	42	61	68.85
Mean	72	100	72	36	64	55	Mean	82	100	82	48	68	73

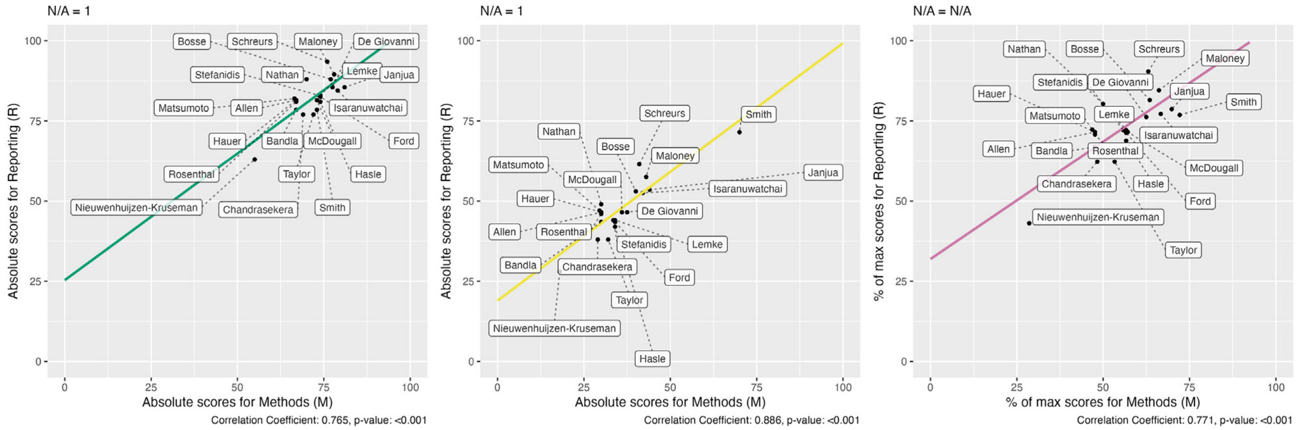
Note: Within each table, we have provided the scores in a scenario where we assign a full score to items that were assessed as N/A (N/A = full score), and a scenario in which we exclude the N/A score (N/A = N/A). The scores in the table are rounded, whereas scores were not rounded to calculate the mean.
CHEQUE indicates checklist for health economic quality evaluations; N/A, not applicable.

ensuring that articles are not penalized for non-applicable items without having to provide both the numerator and denominator. Furthermore, the tool's comprehensiveness also leads to a high burden of reporting on many items, taking the authors on average 20 to 30 minutes per included study to judge 48 items.

In applying the tool, we encountered several ambiguities in interpreting and answering questions. For instance, a question of whether the "best possible option" is considered (M3), implies a binary "yes" or "no" response. A rubric with examples for each question could enhance consistency in judgments across articles and reviewers. In addition, the applicability of certain questions and domains, such as "data inputs and evidence synthesis," to non-modeling studies was unclear. Clarifying which items apply to which analysis types, or whether CHEQUE is intended for use with full or partial economic evaluations, model-based or non-model-based analyses, could address this. Finally, some questions, like M1 (whether the analysis answers an important question for decision-making), could be interpreted on the basis of reviewer judgment, or require specific evidence from the original paper such as stakeholder analysis. Suggestions for specific evidence or examples of questions to consider, without them necessarily becoming separately scored items, could aid reviewers in making judgments and not overlook important areas of consideration.

This ambiguity extends to the deliberate omission of an item on funding sources in CHEQUE, a factor that could significantly impact the reliability of economic evaluations, especially in the presence of conflicts of interest. In our context of medical

Figure 2. Correlation between methods and reporting quality scores in absolute and percentages of maximum score. Left plot: Full scores are awarded to items scored as N/A, and scores are expressed as absolute score (for N/A =1, the maximum obtainable score is 100, and thus the absolute score is the same as the percentage score). Middle plot: N/A-scored items were excluded from the score, and scores are expressed as absolute scores. Right plot: N/A-scored items were excluded from the score, and scores are expressed as the percentage of the maximum attainable score.



N/A indicates not applicable.

education, companies developing high-cost simulators might have a vested interest in portraying their interventions as cost-effective. Whereas CHEQUE's authors argued that biases related to funding sources might be indirectly captured through other quality attributes,[19] this leaves room for interpretation. Suggested specific sources of evidence or triggers within items like M16 (on sources of bias) could remind reviewers to consider funding as a potential source of bias.

Despite its comprehensive nature, scoring 48 different items could be perceived as time-consuming, and the absence of a weighted final judgment might pose challenges for interpretation. Unlike other tools, such as the risk of bias assessments, in which multiple-item questions fall within 1 domain score, CHEQUE's subdomains such as "modeling– M9 to M13" or "data inputs and evidence synthesis –M14 to M16" do not seem to play a specific role in scoring or visualization, and neither does CHEQUE provide an interpretation of the overall score. Subdomains could potentially be assigned a more explicit role either by summarizing multiple items into a smaller number of scores to allow for a faster overview (for example, by using either the average or lowest score of domain items as a domain-score) or by more explicitly stating whether a subdomain applies to all economic evaluations or only specific study types such as decision analytic models to remove some of the scoring ambiguities described in Table 2.

The heterogeneity in study designs, settings, and methodological approaches inherent to health economic evaluations often complicates the process of conducting a meaningful meta-analysis.[1]

Nevertheless, while our review consisted of a qualitative synthesis of included articles, such scores could be of interest if, for example, a meta-analysis would want to weight individual studies in the overall study or exclude studies based on low-quality assessments. Future developments of the tool might benefit from considering the functional role of domains to reduce the number of scores to be interpreted, or recommendations on the interpretation of score ranges.

We presented both the scores in which N/A was assigned a full score and the score in which N/A's were excluded, as we had specified in our protocol. In hindsight, we may have made a deliberate choice to count N/A as a N/A score rather than a full score, as full scores would unintentionally favor the many trials with CEAs above decision analytic models in our review, as all modeling questions were not applicable. In addition, we believed that expressing scores as a percentage, factoring in the maximum attainable score under the N/A equals N/A condition proved more fitting than relying on the absolute scores calculated by the original CHEQUE framework. Using a percentage-based system ensures that articles are not penalized for non-applicable items.

Our study also had limitations. The scope was limited to 1 application within a specific systematic review, with only 2 reviewers assessing the included articles for quality. Our field, medical education, presented unique characteristics that might differ from the more common health outcome-related health economic evaluations. Our review additionally included only 1 model-based study, whereas the CHEQUE questions seem to be primarily aimed at model-based economic evaluations. Finally, because we assessed the studies collaboratively, we were unable to report independent interrater reliability.

Future research directions could explore cross-tool comparisons. This could be across economic evaluation tools using the Grading System for the Quality of Cost-Effectiveness Studies by Chiou et al[3] (as this tool also assesses the reporting aspects), or across fields by comparing the CHEQUE assessment of the economic evaluation to a ROB score of the underlying trial for those alongside trial CEAs. Investigating interrater kappa statistics could provide further insight into the tool's interrater reliability across various types of evaluations. A diverse group of stakeholders, including (systematic review) researchers, patient organizations, health technology organizations, policymakers, and decision modelers, could provide valuable perspectives in further refining and validating the CHEQUE tool. While recognizing that guidance cannot and should not be exhaustive, given the many context-specific situations, additional examples or tutorials could enhance the tool's usability and applicability across diverse settings. The experience gained in our study contributes to the

Table 2. Considerations, examples and dilemmas.

Domain	Items	Examples of considerations
Modeling		
Decision problem and scope	M1. The analysis answers an important question for decision making.	We awarded full scores to all articles. While this question is an essential aspect of study quality, whether or not the question is important depends on context. Had we taken a stricter view we could have required evidence by the authors for a stakeholder need and priority for the research to be conducted.
	M2. The study objective (decision problem) is measurable.	We awarded full scores to all articles. As inclusion criteria were a quantitatively expressed cost and effect, all objectives were measurable. A stricter view could have assessed whether the stated overall study objective was measurable and corresponded to the CEA outcomes.
Intervention and comparator(s)	M3. The comparator(s) is/are the best possible option that appropriately measures the opportunity cost of using the new treatment.	Only one study[39] was assigned a "no" score. The authors set out to compare a PBL curriculum to curricula which had "less" PBL. They did so by comparing student performance and satisfaction in their own university to other universities, without quantifying the extent to which PBL was applied, or potential other differences. As M3 asks for the "best possible" option, this question could be interpreted as only allowing a "yes" or "no" response.
Perspective	M4. The analytic perspective(s) is/are appropriate to answer the research question posed.	Only 1 study[38] which applied a societal perspective, received full score. Remaining studies were graded as "somewhat." The second panel on cost-effectiveness in health and medicine[46] recommends that each CEA should at least contain a health and societal perspective. In the field of education, ideally consequences to the university, student and patients would be incorporated where appropriate. An alternative grading system would have considered the chosen perspectives as still appropriate, however incomplete.
Population	M5. The scope of the study encompasses all populations affected by the intervention.	Similar to item M4, most scores were partial as they did not address consequences to students or patients. Two studies were outliers and graded as "no" as they reviewed the cost-effectiveness of skills trainings in medical students whereas the conclusions were intended for medical residents.[27,33]
Outcome measures	M6. Health outcomes are measured in health metrics that aggregate survival and health-related quality of life or disability (eg, QALY or DALY).	Most studies were graded N/A as their questions did not answer questions on health-related problems, but rather for example the cost-effectiveness of a high-fidelity simulation versus a cardboard structure to teach surgical skills. The choice in assigning full N/A score or excluding N/As influences whether an article is automatically penalized by not including a health-related outcome.
Time horizon	M7. The analytic time horizon is sufficiently long enough to reflect all important differences between intervention(s) and comparator(s).	Only 1 study[38] used a lifetime horizon. All except 3 studies only looked at immediate outcomes which we deemed insufficient to reflect all differences, for example to patient care. However, one could argue that in, for instance, cost-minimization studies the effect outcomes are assumed to be negligible and the immediate cost reduction (assuming no maintenance costs) would be sufficient to reflect all differences.
Discounting	M8. Costs and health effects that occur in the future are discounted to their present value using a recommended discount rate.	As most costs and outcomes were immediate, this item was not applicable in most cases. As the insufficient follow-up was addressed in M7 we decided not to also deduct scores in M8 for this reason alone, however, in such cases an alternative score that could be considered was "no."
Modeling	M9. The chosen model type is appropriate to address study questions.	Only 1 study[38] concerned a modeling study. This study concerned an investigation on the effects of serological testing versus universal hepatitis A vaccination which they investigated through a state-transition cohort model. While a dynamic model or agent-based model allowing interaction between individuals could have been considered more appropriate and the authors make no explicit assumptions on lack of interaction, we still awarded full score as the assumptions needed seemed reasonable.
	M10. The structure of the model reflects the underlying health condition and the impact of the interventions.	We awarded full score for the susceptible, immune, acute hepatitis A infection or dead model structure of our included modeling study.[38]
	M11. Modeling assumptions are reasonable, given the underlying data.	We awarded full score for the explicitly stated assumptions in the one modeling study.[38] However, this judgement requires topic knowledge.
	M12. The need for extrapolation or integrating multiple data sources is considered.	We awarded full score for the one modeling study which used data from multiple sources to inform their model, and extrapolated findings to broader contexts and the future.[38]
	M13. Model validation is conducted, including an assessment of the model structure, assumptions, data, and results.	We awarded a "no" score to the 1 modeling study since it did not conduct model validation. Remaining studies were assigned N/A as this item is a subheading of the modeling domain. Alternatively, we could have demanded some form of verification or validation in other settings.

continued on next page

Table 2. Continued

Domain	Items	Examples of considerations
Data inputs and evidence synthesis	M14. A “best available evidence” approach is used to select data sources for model parameters (eg, conducted or references systematic reviews/meta-analyses).	Most articles received a partial score. These studies used data from their own trials, which we assume in their case could be the best available evidence to their situation, even if authors do not systematically search for other evidence on costs or effects to supplement their economic evaluation. A literal read of this question could also have led to a “no” score if no best available evidence approach was explicitly applied. An alternative interpretation to the questions under subdomain “data inputs and evidence synthesis” (M14-M16) would be to view these strictly as model inputs, in which case all articles except[38] would be scored as N/A.
	M15. Data inputs are generated by appropriate statistical and epidemiological techniques.	We awarded full scores to all but 1 study[39] which compared a universities performance to other universities and assumed the difference was owing to the difference in focus on PBL (as also described in item M3), which we consider to be an inappropriate comparison). However, there is room for nuance in this item. Many studies did not report any uncertainty surrounding their cost outcomes and sometimes their effects. Additionally, we noticed that in the manner in which we filled out the CHEQUE tool, no item distinguished whether data originated from a cohort or RCT, or whether a trial contained 10 or 1000 students. We considered subtracting points under this item owing to the nature of the item but felt that it was not entirely in line with the question asked.
	M16. The quality of the data, including sources of bias, is assessed appropriately.	We expected authors to explicitly address the quality of the data, and in the case of a trial, their own study in light of sources of bias or alternative data resources. Nearly all studies were awarded less than full score.
Consequences	M17. Major consequences affected by the choice of interventions being compared are identified.	We awarded nearly all partial scores as studies looked at direct and immediate consequences of interventions such as an increase in student performance but neglected to consider other consequences to students (for example, their well-being or time spent in practicing the intervention), faculty or patients. We considered whether we were deducting scores for the same issue twice within studies that did not adopt a broader societal perspective under item M4 but decided that even from a payer/university perspective we would expect consequences to faculty to be included.
Utilities (preference measures)	M18. Health preferences reflect those of the jurisdiction(s) of interest (as specified in the decision problem).	Only 1[38] study reported QALYs as an outcome. They assumed the quality of life in the hepatitis A state was 11/12th of a year, as 1 month of work would be missed. We did not consider this a reflection of the population’s preferences.
Costs and resource use	M19. Resource use that is nontrivial in magnitude is included in the reference case analysis.	We awarded most studies that investigated the immediate costs with full scores. In a study considering a high-cost heart simulator the study included the purchase cost but not maintenance cost in their assessment, this study received a partial score. One could interpret this question as only answerable with yes/no, in this case the awarded score would be “no.”
Analysis	M20. Incremental analyses are conducted (i.e., the additional costs generated by one alternative over another are compared with the additional effects generated).	Approximately half of included studies provided information on incremental costs and effects, the remaining studies only report the costs and effects for each group. It was unclear in which situations a score of “somewhat” would be assigned. This could have been done if only incremental costs or only incremental effects were reported, however, we still assigned a “no” in such cases.
	M21. ICERs are obtained by comparing each intervention with the next most effective option after eliminating dominated options.	While the question asks to eliminate dominated options, it was not immediately clear whether we should assign full scores or N/A scores in studies which had dominated options but did not explicitly identify them as such. We scored these as N/A.
	M22. Probabilistic sensitivity analysis is conducted to account for uncertainty in input parameters simultaneously.	The modeling study[38] in our review did not include a probabilistic sensitivity analysis and was awarded a “no” score. Two other studies used bootstrapping to construct a PSA[21,32] and cost-effectiveness acceptability curve and were awarded a “yes” score.
	M23. Alternative modeling choices and assumptions (structural uncertainty) are explored through additional sensitivity analysis (ie, scenario analysis).	We interpreted this question more broadly beyond only “models” by reviewing whether any alternative assumptions or scenarios were explored. Only a few studies did so and received partial or full scores. For example, one study[22] investigated what happened to the ICER if staff time was increased or if IT support was added.
Equity considerations	M24. Relevant equity or distributional considerations are taken into account.	No study discussed equity or distributional considerations. This question in CHEQUE helps highlight the fact that these considerations are too often overlooked. Especially in areas of student assessment and selection and the involvement of patient educators, equity concerns have become more prominent in the overall discourse of the field, but these discussions were not reflected in CEAs on these topics.

continued on next page

Table 2. Continued

Domain	Items	Examples of considerations
Reporting		
Decision problem and scope	R1. The study objectives (or decision problems) are clearly stated.	The study objectives were clearly stated in all articles.
Intervention and comparator(s)	R2. All aspects of the interventions that may affect their cost-effectiveness are clearly defined (eg, frequency of delivery, setting of delivery, and specific technologies used). R3. The comparator(s) is/are clearly stated.	Nearly all articles clearly defined the interventions, except a study that set out to identify the effect of PBL[39] but did not quantify the differences between the intervention and comparator groups in PBL exposure or other factors that could influence the outcomes (previously described in item M3 and M15). This article received a “somewhat” score. Alternatively, if one were to consider this as a “yes or no” question, since either “all aspects” are or are not defined, we could have considered awarding a “no” instead. A study on the effect of PBL[39] was assigned a “somewhat” score as no information was provided on the quantity of intervention exposure in the comparator, or other differences with the intervention group that could cause the outcome (as also listed in items R2, M3, and M15).
Perspective	R4. The analytic perspective(s) is/are clearly stated.	Only 2 studies[26,38] explicitly stated the perspective taken, the other studies were awarded a “no”. However, from the information in the paper we could infer that these were all from a payer (university) perspective.
Population	R5. The target population is clearly stated.	We awarded all articles full scores.
Outcome measures	R6. Primary outcome measures are clearly stated. R7. ICERs are reported.	The outcome measures were stated in all articles, but not always clearly defined as primary outcome. If there was only 1 outcome in the article, we still awarded full scores if this was the case. If multiple outcomes were well defined but no primary outcome was assigned, we assigned a partial score. The item could be interpreted as requiring a clear definition of outcome measures, or the primary outcome identification specifically. We assigned full score if ICERs were reported. If there was a dominant situation but this was not explicitly reported as one, we assigned an “N/A” score
Time Horizon	R8. The analytic time horizon is clearly stated.	In most cases no time horizon was clearly specified. One study[32] only did so in the supplementary materials, and because the question asks whether the time horizon was “clearly” stated, was still awarded a “no” score. Alternatively, as the horizon is still stated in materials related to the paper, we could have also awarded a “yes”.
Discounting	R9. The discount rate is clearly stated.	In most cases, the discount rate was not clearly specified. One study[32] only did so in the supplementary materials and because the question asks whether the rate was “clearly” stated was still awarded a “no” score. Alternatively, as the discount rate is still stated in materials related to the paper, we could have also awarded a “yes.”
Modeling	R10. The type of model used is clearly stated. R11. Justification of modeling choices and assumptions is provided. R12. Model descriptions are detailed enough to allow for replication. R13. The description of how the model was validated is provided. R14. The software used to develop the model is clearly stated.	The modelling study[38] clearly identified itself as a Markov-Model. The modelling study[38] provided justification for some modeling choices but not all and received a “somewhat” score. Sufficient information was provided to allow for replication of the modelling study.[38] As no article performed validation, all received an N/A score. The software was not clearly stated in the modelling study. While the item falls under the modelling domain, we would argue that clearly stating software used for any type of analysis—including non-model-based studies—is relevant. Nevertheless, as the item falls under modelling the remaining studies received an N/A score.
Data inputs and evidence synthesis	R15. All data sources are clearly referenced.	All articles except 1 (Nieuwehuijzen-Kruseman)[39] stated their data sources.
Consequences	R16. Comprehensive identification of potential consequences is summarized (eg, using an Impact Inventory table in Second Panel’s report).	No paper created an explicit impact inventory of all consequences, including those outside of the formal healthcare sector/perspective.
Utilities (preference measures)	R17. Sources for the utility weights are clearly stated.	The study using utility weights[38] does provide a reference. Nevertheless, it is unclear why the authors use a source that states health workers lose 1 month of work if in the hepatitis A state as a source for a utility weight. We therefore assigned a “somewhat” score.

continued on next page

Table 2. Continued

Domain	Items	Examples of considerations
Costs and resource use	R18. Quantities of resources are reported separately from the prices (unit costs) of those resources.	The majority of studies reported quantities of resources separate from prices. Some only provided a lump sum, for example, a study on laparoscopic skill training[24] stated that the conventional simulation training was 30 000 € more expensive than the cardboard box, without further specification.
Analysis	R19. The approach to secondary analyses (eg, sensitivity, scenario, or subgroup analysis) is sufficiently described.	The majority of articles conducted no secondary analyses. When no such analyses were conducted, we scored reporting as N/A. Some articles provided a partial explanation and partial results of a sensitivity analysis and received a "somewhat" score. For example, a study on the selection of students discusses a supplemental hypothetical additional cohort of students and how this would affect cost-benefits, but do not provide further details, and do not describe sensitivity analyses in the methods section.[26]
Equity considerations	R20. Discussion section includes a description of any significant ethical implications of the CEA results.	While none of the articles took ethical considerations into account in their analyses (item M24), we still assigned a "no" score to all articles. This could be considered deducting points twice for the same issue by assigning "no" instead of N/A, however, even if the analysis itself does not consider ethical implications in the methodology, discussion can still address potential considerations for the future.
Transparency and reporting	R21. Results are presented in a disaggregated format for transparency.	Most studies received full scores. We assigned partial scores if for example effects were described for each intervention and incrementally, but costs were only described incrementally.[24] The phrasing of the item however does leave room for interpretation on what is required to be considered disaggregated.
	R22. The relevance of study results to specific decision problems is discussed.	All articles discussed the relevance of their findings.
	R23. Implications of uncertainty for decision making, including the need for future research, are explored.	Most articles discussed areas for future research, although not all did. However, many studies did not incorporate uncertainty in cost-effectiveness, and not all specifically related to uncertainties in decision making. If they did not reflect on the uncertainty of their conclusion, they were awarded a "no" score.
	R24. Potential bias and limitations are discussed.	Most articles did discuss limitations and to a lesser extent specific forms of bias. One study did discuss bias but concluded that there was none because it was a randomized controlled trial. We awarded this paper with a "somewhat" score, while they technically do discuss bias, they did not capture major sources of bias in their study. As the CHEQUE tool does not specifically ask for industry sponsorship[46] and conflicts of interests, we considered that this item might still reflect part of such bias. Nevertheless, no conflicts of interest of this nature were listed in our included articles.

CEA indicates cost-effective analysis; CHEQUE indicates checklist for health economic quality evaluations; DALY, disability-adjusted life year; ICER, incremental cost-effectiveness ratio; IT, information technology; N/A, not applicable; PBL, problem-based-learning; PSA, probabilistic sensitivity analyses; QALY, quality-adjusted life year; RCT, randomized control studies.

ongoing dialogue on refining and advancing tools for quality assessment in economic evaluations.

Conclusion

Our application of the CHEQUE tool in a systematic review setting on economic evaluations in medical education highlights its adaptability and comprehensive nature. While acknowledging its strengths, such as systematic and standardized assessment, our insights suggest potential areas for refinement, including scoring ambiguities and scoring burdens that need addressing. We foresee ongoing collaboration and improvements in health economic research methodology, aiming to enhance the utility of CHEQUE across diverse applications. This article contributes to the dialogue on refining assessment tools, urging continued development to meet the evolving needs of the field.

Author Disclosures

Author disclosure forms can be accessed below in the Supplemental Material section.

Supplemental Material

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2024.10.3853.

Article and Author Information

Accepted for Publication: October 8, 2024

Published Online: January 3, 2025

doi: https://doi.org/10.1016/j.jval.2024.10.3853

Author Affiliations: Department of Epidemiology (Dijk, Hunink), and Department of Radiology, Erasmus MC, University Medical Center Rotterdam, The Netherlands (Dijk, Hunink); Department of Gastroenterology and Hepatology, HagaZiekenhuis, The Hague, Zuid-Holland, The Netherlands (Dijk); Department of Radiology, Elisabeth-Tweesteden Ziekenhuis, Tilburg, Noord-Brabant, The Netherlands (Dijk); Erasmus School of Health Policy & Management, Erasmus University Rotterdam, Zuid-Holland, Rotterdam, The Netherlands (Essafi); Center for Health Decision Science, Harvard T.H. Chan School of Public Health, Boston, Massachusetts (Hunink).

Correspondence: Myriam G.M. Hunink, MD, PhD, Professor of Clinical Epidemiology and Radiology, Department of Epidemiology, Office NA 28-18, Erasmus Medical Centre, PO Box 2040, Zuid-Holland, Rotterdam 3000CA, The Netherlands. Email: m.hunink@erasmusmc.nl

Authorship Confirmation: All authors certify that they meet the ICMJE criteria for authorship.

Funding/Support: The authors received no financial support for this research.

Acknowledgments: The authors thank the authors of the CHEQUE tool for their efforts in creating a comprehensive tool that can assess the quality of methods and reporting in health economic literature. The authors also thank Christa Niehot, Andrea Woltman, and John Wong for their contributions to the Systematic Review that we used as an example in this paper.

Protocol Registration: The protocol for the systematic review was preregistered in PROSPERO under registration number CRD42023478907. Amendments made after title abstract screening commenced but before full-text screening include a specification of a minimum percentage of medical students in the population and language requirement.

Data Availability: All data and R code used to generate scores and figures are provided [https://github.com/stijntje-dijk/CHEQUE]

REFERENCES

1. Shields GE, Elvidge J. Challenges in synthesising cost-effectiveness estimates. *Syst Rev*. 2020;9(1):289.
2. Walker DG, Wilson RF, Sharma R, et al. *Best Practices For Conducting Economic Evaluations In Health Care: A Systematic Review Of Quality Assessment Tools*. Rockville, MD: Agency for Healthcare Research and Quality; 2012.
3. Chiou C-F, Hay JW, Wallace JF, et al. Development and validation of a grading system for the quality of cost-effectiveness studies. *Med Care*. 2003;41(1):32–44.
4. Ungar WJ, Santos MT. The Pediatric Quality Appraisal Questionnaire: an instrument for evaluation of the pediatric health economics literature. *Value Heal*. 2003;6(5):584–594.
5. Adams ME, McCall NT, Gray DT, Orza MJ, Chalmers TC. Economic analysis in randomized control trials. *Med Care*. 1992;30(3):231–243.
6. Drummond ME, Sculpher MJ, Torrance GW, O'Brien BJ, Stoddart GL. Critical assessment of economic evaluation. In: Drummond ME, Sculpher MJ, Torrance GW, O'Brien BJ, Stoddart GL, Drummond ME, Sculpher MJ, Torrance GW, O'Brien BJ, Stoddart GL, eds. *Methods for the Economic Evaluation of Health Care Programme*. 3rd ed. Oxford University Press; 2005:27–54.
7. Gerard K. Cost-utility in practice: a policy maker's guide to the state of the art. *Health Policy (New York)*. 1992;21(3):249–279.
8. Sacristán JA, Soto J, Galende I. Evaluation of pharmacoeconomic studies: utilization of a checklist. *Ann Pharmacother*. 1993;27(9):1126–1133.
9. Clemens K, Townsend R, Luscombe F, Mauskopf J, Osterhaus J, Bobula J. Methodological and conduct principles for pharmacoeconomic research. *Pharmacoeconomics*. 1995;8(2):169–174.
10. Russell LB, Gold MR, Siegel JE, Daniels N, Weinstein MC. The role of cost-effectiveness analysis in health and medicine. *JAMA*. 1996;276(14):1172–1177.
11. Sanders GD, Neumann PJ, Basu A, et al. Recommendations for conduct, methodological practices, and reporting of cost-effectiveness analyses: second panel on cost-effectiveness in health and medicine. *JAMA*. 2016;316(10):1093–1103.
12. Drummond MF, Jefferson TO. Guidelines for authors and peer reviewers of economic submissions to the BMJ. The BMJ Economic Evaluation Working Party. *Br Med J*. 1996;313(7052):275.
13. Grutters JPC, Seferina SC, Tjan-Heijnen VCG, van Kampen RJW, Goettsch WG, Joore MA. Bridging trial and decision: a checklist to frame health technology assessments for resource allocation decisions. *Value Heal*. 2011;14(5):777–784.
14. Kunst N, Siu A, Drummond M, et al. CHEERS value of information (CHEERS-VOI) reporting standards–explanation and elaboration. *Value Heal*. 2023;26(10):1461–1473.
15. Husereau D, Drummond M, Augustovski F, et al. Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 2022) statement: updated reporting guidance for health economic evaluations. *Int J Technol Assess Health Care*. 2022;38(1):e13.
16. Husereau D, Drummond M, Petrou S, et al. Consolidated health economic evaluation reporting standards (CHEERS) statement. *Eur J Heal Econ*. 2013;14(3):367–372.
17. Philips Z, Bojke L, Sculpher M, Claxton K, Golder S. Good practice guidelines for decision-analytic modelling in health technology assessment: a review and consolidation of quality assessment. *Pharmacoeconomics*. 2006;24(4):355–371.
18. Noben CY, de Rijk A, Nijhuis F, Kottner J, Evers S. The exchangeability of self-reports and administrative health care resource use measurements: assessement of the methodological reporting quality. *J Clin Epidemiol*. 2016;74:93–106.
19. Kim DD, Do LA, Synnott PG, et al. Developing criteria for health economic quality evaluation tool. *Value Heal*. 2023;26(8):1225–1234.
20. Imbens GW. Potential outcome and directed acyclic graph approaches to causality: relevance for empirical practice in economics. *J Econ Lit*. 2020;58(4):1129–1179.
21. Isaranuwatchai W, Brydges R, Carnahan H, Backstein D, Dubrowski A. Comparing the cost-effectiveness of simulation modalities: a case study of peripheral intravenous catheterization training. *Adv Heal Sci Educ*. 2014;19(2):219–232.
22. Maloney S, Nicklen P, Rivers G, et al. A cost-effectiveness analysis of blended versus face-to-face delivery of evidence-based medicine to medical students. *J Med Internet Res*. 2015;17(7):e4346.
23. Bandla H, Franco RA, Simpson D, Brennan K, McKanry J, Bragg D. Assessing learning outcomes and cost effectiveness of an online sleep curriculum for medical students. *J Clin Sleep Med*. 2012;8(4):439–443.
24. Chandrasekera SK, Donohue JF, Orley D, et al. Basic laparoscopic surgical training: examination of a low-cost alternative. *Eur Urol*. 2006;50(6):1285–1291.
25. Bosse HM, Nickel M, Huwendiek S, Schultz JH, Nikendei C. Cost-effectiveness of peer role play and standardized patients in undergraduate communication training Approaches to teaching and learning. *BMC Med Educ*. 2015;15(1):4–9.
26. Schreurs S, Cleland J, Muijtjens AMM, Oude Egbrink MGA, Cleutjens K. Does selection pay off? A cost–benefit comparison of medical school selection and lottery systems. *Med Educ*. 2018;52(12):1240–1248.
27. Stefanidis D, Hope WW, Korndorffer Jr JR, Markley S, Scott DJ. Initial laparoscopic basic skills training shortens the learning curve of laparoscopic suturing and is cost-effective. *J Am Coll Surg*. 2010;210(4):436–440.
28. Lemke M, Lia H, Gabinet-Equihua A, et al. Optimizing resource utilization during proficiency-based training of suturing skills in medical students: a randomized controlled trial of faculty-led, peer tutor-led, and holography-augmented methods of teaching. *Surg Endosc*. 2020;34(4):1678–1687.
29. Taylor CA, Green KE. OSCE feedback: a randomized trial of effectiveness, cost-effectiveness and student satisfaction. *Creat Educ*. 2013;4(06):9.
30. Rosenthal ME, Castellvi AO, Goova MT, Hollett LA, Dale J, Scott DJ. Pretraining on Southwestern stations decreases training time and cost for proficiency-based fundamentals of laparoscopic surgery training. *J Am Coll Surg*. 2009;209(5):626–631.
31. Ford H, Cleland J, Thomas I. Simulated ward round: reducing costs, not outcomes. *Clin Teach*. 2017;14(1):49–54.
32. Janjua A, Roberts T, Okeahialam N, Clark TJ. Cost-effective analysis of teaching pelvic examination skills using Gynaecology Teaching Associates (GTAs) compared with manikin models (The CEAT Study). *BMJ Open*. 2018;8(6):e015823.
33. Matsumoto ED, Hamstra SJ, Radomski SB, Cusimano MD. The effect of bench model fidelity on endourological skills: a randomized controlled study. *J Urol*. 2002;167(3):1243–1247.
34. Nathan A, Fricker M, Georgi M, et al. Virtual interactive surgical skills classroom: a parallel-group, non-inferiority, adjudicator-blinded, randomised controlled trial (VIRTUAL). *J Surg Educ*. 2022;79(3):791–801.
35. Hauer KE, Chou CL, Souza KH, et al. Impact of an in-person versus web-based practice standardized patient examination on student performance on a subsequent high-stakes standardized patient examination. *Teach Learn Med*. 2009;21(4):284–290.
36. McDougall EM, Kolla SB, Santos RT, et al. Preliminary study of virtual reality and model simulation for learning laparoscopic suturing skills. *J Urol*. 2009;182(3):1018–1025.
37. De Giovanni D, Roberts T, Norman G. Relative effectiveness of high-versus low-fidelity simulation in learning heart sounds. *Med Educ*. 2009;43(7):661–668.
38. Smith S, Weber S, Wiblin T, Nettleman M. Cost-effectiveness of hepatitis A vaccination in healthcare workers. *Infect Control Hosp Epidemiol*. 1997;18(10):688–691.
39. Nieuwenhuijzen Kruseman A, Kolle LFJTM, Scherpbier AJJ. Problem-based learning at Maastricht. An assessment of cost and outcome. *Educ Heal*. 1997;10:179–187.
40. Hasle JL, Anderson DS, Szerlip HM. Analysis of the costs and benefits of using standardized patients to help teach physical diagnosis. *Acad Med*. 1994;69(7):567–570.
41. Allen SS, Miller J, Ratner E, Santilli J. The educational and financial impact of using patient educators to teach introductory physical exam skills. *Med Teach*. 2011;33(11):911–918.
42. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ*. 2019;366:l4898.
43. Sterne JAC, Hernán MA, Reeves BC, et al. Robins-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*. 2016;355:i4919.
44. R Core team. R: A Language and Environment for Statistical Computing. https://www.r-project.org/; Published 2013.
45. McGuinness LA. robvis: an R package and web application for visualising risk-of-bias assessments. GitHub URL. https//githubcom/mcguinlu/robvis; Published 2019. Accessed July 31, 2021.
46. Bell CM, Urbach DR, Ray JG, et al. Bias in published cost effectiveness studies: systematic review. *BMJ*. 2006;332(7543):699–703.