Canadian Medical Education Journal

Original Research

'*What would my peers say?*' Comparing the opinion-based method with the prediction-based method in Continuing Medical Education course evaluation

Qu'en penseraient mes pairs? Comparaison entre la méthode fondée sur l'opinion et celle fondée sur la prédiction dans l'évaluation de cours de formation médicale continue

Jamie S Chua,[1] Merel van Diepen,[2] Marjolijn D Trietsch,[3] Friedo W Dekker,[2,4] Johanna Schönrock-Adema,[5 6] Jacqueline Bustraan[4]

[1]Department of Gastroenterology and Hepatology, Leiden University Medical Center, Leiden, The Netherlands; [2]Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands; [3]Department of Gynaecology, Leiden University Medical Center, Leiden, The Netherlands; [4]Center for Innovation in Medical Education, Department of Education, Leiden University Medical Center, Leiden, The Netherlands; [5]Center for Educational Development and Research in health sciences, University of Groningen and University Medical Center Groningen, Groningen, The Netherlands; [6]Hanze University of Applied Sciences, Groningen, The Netherlands

Correspondence to: Jamie S. Chua, MD, PhD; Department of Gastroenterology and Hepatology, Leiden University Medical Center, Albinusdreef 2, C4P, 2300 RC, Leiden, The Netherlands; phone: +31-(0)71-526 5008; email: j.s.chua@lumc.nl

Published ahead of issue: Apr 22, 2024; published: Jul 12, 2024; CMEJ 2024 Available at https://doi.org/10.36834/cmej.77580

Abstract

Background: Although medical courses are frequently evaluated via surveys with Likert scales ranging from "*strongly agree*" to "*strongly disagree,*" low response rates limit their utility. In undergraduate medical education, a new method with students predicting what their peers would say, required fewer respondents to obtain similar results. However, this prediction-based method lacks validation for continuing medical education (CME), which typically targets a more heterogeneous group than medical students.

Methods: In this study, 597 participants of a large CME course were randomly assigned to either express personal opinions on a five-point Likert scale (opinion-based method; *n* = 300) or to predict the percentage of their peers choosing each Likert scale option (prediction-based method; *n* = 297). For each question, we calculated the minimum numbers of respondents needed for stable average results using an iterative algorithm. We compared mean scores and the distribution of scores between both methods.

Results: The overall response rate was 47%. The prediction-based method required fewer respondents than the opinion-based method for similar average responses. Mean response scores were similar in both groups for most questions, but prediction-based outcomes resulted in fewer extreme responses (strongly agree/disagree).

Conclusions: We validated the prediction-based method in evaluating CME. We also provide practical considerations for applying this method.

Résumé

Contexte : Bien que les cours de médecine soient fréquemment évalués au moyen d'enquêtes avec des échelles de Likert allant de « totalement d'accord » à « totalement en désaccord », les faibles taux de réponse en limitent l'utilité. Dans l'enseignement médical prédoctoral, une nouvelle méthode dans laquelle les étudiants prédisent ce que leurs pairs diraient, nécessite moins de répondants pour obtenir des résultats similaires. Cependant, cette méthode fondée sur la prédiction n'est pas validée pour la formation médicale continue (FMC), qui cible généralement un groupe plus hétérogène que les étudiants en médecine.

Méthodes : Dans cette étude, 597 participants à un grand cours de FMC ont été choisis au hasard pour exprimer leur opinion personnelle sur une échelle de Likert en cinq points (méthode fondée sur l'opinion; n = 300) ou à prédire le pourcentage de leurs pairs choisissant chaque option de l'échelle de Likert (méthode fondée sur la prédiction; n = 297). Pour chaque question, nous avons calculé le nombre minimum de répondants nécessaire pour obtenir des résultats moyens stables à l'aide d'un algorithme itératif. Nous avons comparé les scores moyens et la distribution des scores entre les deux méthodes.

Résultats : Le taux de réponse global était de 47 %. La méthode fondée sur la prédiction a nécessité moins de répondants que celle fondée sur l'opinion pour des réponses moyennes similaires. Les scores moyens des réponses étaient similaires dans les deux groupes pour la plupart des questions, mais les résultats fondés sur la prédiction ont donné lieu à moins de réponses extrêmes (totalement d'accord/totalement en désaccord).

Conclusions : Nous avons validé la méthode fondée sur la prédiction dans l'évaluation de la FMC. Nous présentons également des considérations pratiques pour la mise en œuvre de cette méthode.

Introduction

Providing high quality education is one of the main responsibilities of medical schools. Surveys are important tools to monitor and improve the quality of education. Response rates of at least 70% have been considered necessary to ensure valid survey results and to minimize response bias, as lower response rates may not adequately represent the views of the larger population being surveyed.[1-3] However, in general, response rates are low.[4,5]

In Continuing Medical Education (CME) targeting physicians, response rates of surveys are often even lower, with some rates below 5%.[2,6-9] Potential explanations for these low response rates include 'evaluation fatigue,' doubts about whether the outcomes will be used, the heavy clinical workload that may prevent respondents to create time and motivation to respond to surveys, lack of interest, limited benefits for those who will not attend the course again, and low perceived value of the evaluation.[8,10,11] When response rates are low, providers of medical education may be hesitant to use evaluation results for optimizing their programmes since reliability decreases with lower response rates.[8]

A promising evaluation method that may help to overcome the problem of low response rates and to improve the validity and utility evaluation outcomes is the prediction-based method. This method is derived from the field of political science where respondents were asked to predict the outcomes of elections.[12,13] In that field, the prediction-based method was demonstrated to require much smaller sample sizes, while resulting in more accurate predictions than traditional opinion polls asking respondents' personal opinions, and this was true even if the specific respondent sample was drawn from a region that clearly differed from the general political preference in terms of political affiliation.[12,13] The rationale behind the prediction approach is that it reduces confounding influences or bias due to personal factors such as prejudices, emotions and personally irrelevant thoughts, and–thus–provides more valid outcomes.[13] Hofstee emanated from Spearman and Brown's axioms concerning True Score Theory when formulating this rationale: he reasoned that averaged predictions of outcomes would lead to better outcomes than averaged opinions. Inspired by this work, Cohen-Schotanus et al. and Schönrock-Adema et al. applied the prediction-based method in course evaluations among undergraduate medical students.[14,15] They discovered that also in undergraduate medical education, the prediction-based method required fewer respondents to obtain similar results and was less susceptible to response bias compared with the opinion-based method which required students to express personal opinions on a Likert scale.[14,15] Recently, these findings were validated in two other studies among undergraduate year 1-3 medical and undergraduate accounting students.[16,17]

The prediction-based method could be helpful to solve the problem of inaccurate and invalid evaluation results due to low response rates in CME. However, despite a clear theoretical rationale behind the prediction-based method and empirical evidence for this method within the undergraduate medical education context, it is by no means clear whether the prediction-based method is valid in the context of CME. Validation in CME is important because CME participants may have less sense of the prevailing opinion than their undergraduate counterparts do. Whereas undergraduate students may share multiple courses and educational experiences over a relatively long period, may spend more time together discussing their opinions – both on and off-campus – and, hence, may be well-acquainted with their peers and their peers' opinions about educational aspects, CME participants often only attend a single course and often meet their peers only briefly during the course. Moreover, the CME population is often a heterogeneous group of participants, coming from different work contexts and varying in level of knowledge and years of professional experience, unlike the undergraduate medical student population, which is relatively homogeneous in terms of age group and education experience. For CME participants in large, heterogeneous groups, it may therefore be difficult to predict how their peers value the quality of the CME course.

Therefore, the aim of this study was to validate the prediction-based method in the distinct setting of CME. We investigated if the prediction-based method yielded outcomes comparable to those obtained with the traditionally used, opinion-based method in the context of a large-scale CME course. By applying the bootstrap method to estimate statistical uncertainty, we aimed to advance the statistical methodology used for prediction-based course evaluation.

Methods

Respondents and procedure

The participants in this study (*n* = 597) were enrolled attendants of the Boerhaave Continuing Medical Education course 'Progress in Practice' at the Leiden University Medical Center. This two-day course aims to update participants on scientific developments in sessions on important domains and patient problems in the general

practice of primary care physicians, for example pharmacogenetics, telemonitoring in cardiology or point-of-care testing in general practice. The course has a long tradition with renowned experts as teachers. Participants were invited to evaluate the course using evaluation forms that were distributed digitally on the last day of the course. We informed the course participants at the beginning of the course about the aim and design of the study, about the estimated time needed to complete the evaluation (approximately 10 minutes), and that participation was on a voluntary basis and anonymous. The work was carried out in accordance with the Declaration of Helsinki and the study was approved by the LUMC Educational Research Review Board, OEC/ERRB/20180817/1.

Course evaluation form

Educationalists developed the course evaluation form at the University of Leiden and the Course Committee and Educational Board reviewed and approved this form. The form consisted of 19 questions (Supplemental Table S1): six multiple choice questions addressing respondent background, 10 questions–all positive statements–evaluating specific course components on a five-point Likert-type scale ranging from "strongly disagree" to "strongly agree," one question evaluating the method of evaluation, two open text questions asking to identify main 'take home messages' and suggestions for improvement of the course.

Study design

We asked half of the participants to complete the evaluation according to the traditional opinion-based method and the other half to complete the evaluation according to the prediction-based method. We assigned the evaluation forms randomly to participants. Participants assigned to the opinion-based method completed the forms as usual by giving their opinions on a list of statements about specific course components.[18,19] Participants assigned to the prediction-based method estimated for each answer option of each question, which percentages of their peers would choose that answer options (see Figure 1 for an example). Per question, the percentages for all response options should add up to 100%.

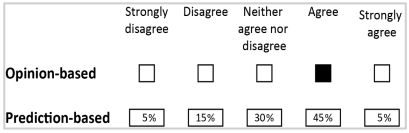

Figure 1. Example of evaluation using the opinion-based and the prediction-based method.[a]

Analysis and statistics

We measured respondent characteristics categorically, described them as numbers (*n*) with percentages (%) and compared them using a Chi-square test. We excluded responses in the prediction-group from the analyses if they did not add up to 100%. To be able to compare the outcomes of the prediction-based and opinion-based methods, we converted the percentages of the respondents in the prediction-group into scores: for each respondent, we calculated the weighted mean response per question by multiplying each Likert scale option (1 to 5) by the weights (percentages) assigned by the respondents to each of the respective response options and summing these values. For the opinion-based method, the weight of their chosen response was one, and the weights of the other Likert scale options were zero, so that their score was equal to the chosen Likert scale option. Next, using a student's *t*-test and MANOVA, we investigated for each question whether the mean scores differed between the two methods. Although it is an assumption that successive response categories on a Likert scale are equidistant, using these parametric methods to analyse Likert scale data have been shown to be tenable as parametric statistics have consistently been shown to be robust to violations of assumptions underlying the use of parametric tests.[20] As the sample sizes in our two groups are sufficiently large, we thus used parametric tests even though our data may not be normally distributed.[21] Since the percentages of the high and low Likert scores–combinations of scores in the highest and lowest response categories respectively–are also frequently used in daily practice to determine the success of a course, we also compared both methods on these aggregated scores. Consequently, we calculated the mean percentages of high and low scores between the two groups, with high scores defined as a score of 4 or 5 and low scores as a score of 1 or 2. We compared the percentages using a Chi-square test.

For each evaluation method, we determined how many respondents were at least required for reliable outcomes, based on a previously published iterative algorithm.[14, 15] This method entailed that we iteratively compared the average outcomes of subsamples in terms of percentages per response option to those of the entire sample by taking a random subsample, existing of one single respondent, for each item and adding one respondent at a time to the subsample in each next iteration, until all respondents were included in the sample. For each subsample, we calculated per question the average percentages per

[a] In the opinion-based method, participants were asked to give their own opinions for each item, by selecting per item the answer option on the Likert scale ranging from "strongly disagree" to "strongly agree." In the prediction-based method, participants were asked to predict which percentages of their peers would choose the different answer options of the five-level Likert scale; for each item, the percentages noted for all the answer options should add up to 100%.

response option. We compared these percentages with those of the entire sample: per subsample per question, we summed the (absolute) differences between the subsample's and entire sample's percentages for each of the five response options. The point in the iterative process where the differences between subsample and total sample were smaller than 5% and did not exceed 5% anymore represented the minimum number of respondents needed for stable outcomes (for more details see Cohen-Schotanus et al.[14] and Schönrock-Adema et al.[15]). To strengthen the validity of our outcomes, we estimated the statistical uncertainty around the minimum respondent numbers by using bootstrap analyses with 1000 repetitions for each of the 10 questions in each of the two evaluation methods. Using random sampling with replacement, we drew a random sample of the original sample size. Next, we ordered this sample randomly, and computed the minimum required number of respondents in this bootstrap sample according to the above-described iterative process. The code is added to the supplemental data. We present means over bootstrap samples per question per condition. Using the percentile method, 95%-confidence intervals were constructed for the differences between conditions in minimum required numbers of respondents in the 1000 bootstrap samples for each question and model. P-values were computed according to the method by Altman and Bland.[22] Briefly, we inferred the standard error from the confidence interval, then computed the Z-statistic and calculated the p-value according to the formula $p = \exp(-0.77*Z - 0.416*Z^2)$. We used the SPSS statistical software package (version 20.0; IBM Corp) and R version 3.2.2 (2015-08-14) -- "Fire Safety" Copyright (C) 2015 The R Foundation for Statistical Computing. Differences with $p < 0.05$ were considered to be statistically significant.

Results

Participant characteristics

The vast majority of the attendants of the course (575 out of 597, i.e., >96%) were general practitioners; the others were elderly Care Physicians, Intellectual Disability Physicians, Physician Assistants and Specialized nurses. The overall response rate was 282/597 (47%). The response rates in the prediction and the opinion-based methods differed significantly with 122/297 (41%) and 160/300 (53%), respectively ($p < 0.01$). Respondent characteristics showed significant differences between the years of experience, times participated, age categories, and gender between respondents in the opinion-based and prediction-based methods (Table 1). In the prediction-based method, on average 12% (between 9% and 16% across questions) of the responses were excluded from analyses as their responses did not add up to 100%.

Table 1. Respondent characteristics

Topic of question	Total *n* = 282	Opinion-based *n* = 160	Prediction-based *n* = 122	p-value
Profession, n (%)				
General practitioner	226 (80.1)	140 (87.5)	86 (70.5)	0.604
General practitioner resident	3 (1.1)	1 (0.6)	2 (1.6)	
Intellectual disability physician	15 (5.3)	9 (5.6)	6 (4.9)	
Intellectual disability physician resident	1 (0.4)	0 (0.0)	1 (0.8)	
Physician assistant	5 (1.8)	3 (1.9)	2 (1.6)	
Nurse specialist	9 (3.2)	4 (2.5)	5 (4.1)	
No response	23 (8.2)	3 (1.9)	20 (16.4)	
Years of experience, n (%)				
0 to 3	24 (8.5)	11 (6.9)	13 (10.7)	0.002
3 to 10	49 (17.4)	18 (11.2)	31 (25.4)	
10 to 25	111 (39.4)	65 (40.6)	46 (37.7)	
> 25	97 (34.4)	66 (41.2)	31 (25.4)	
No response	1 (0.4)	0 (0.0)	1 (0.8)	
Times participated, n (%)				
This is my first time	51 (18.1)	20 (12.5)	31 (25.4)	0.003
2 to 5 times, including this time	87 (30.9)	45 (28.1)	42 (34.4)	
> 5 times	143 (50.7)	94 (58.8)	49 (40.2)	
No response	1 (0.4)	1 (0.6)	0 (0.0)	
Age, n (%)				
25-35 years	28 (9.9)	10 (6.2)	18 (14.8)	0.006
35-45 years	45 (16.0)	20 (12.5)	25 (20.5)	
45-55 years	82 (29.1)	46 (28.8)	36 (29.5)	
>55 years	125 (44.3)	83 (51.9)	42 (34.4)	
No response	2 (0.7)	1 (0.6)	1 (0.8)	
Sex, n (%)				
Male	152 (53.9)	96 (60.0)	56 (45.9)	0.027
Female	126 (44.7)	63 (39.4)	63 (51.6)	
No response	4 (1.4)	1 (0.6)	3 (2.5)	
Region of practice, n (%)				
North	98 (34.8)	55 (34.4)	43 (35.2)	0.821
East	21 (7.4)	13 (8.1)	8 (6.6)	
Middle	21 (7.4)	14 (8.8)	7 (5.7)	
West	81 (28.7)	43 (26.9)	38 (31.1)	
South	59 (20.9)	33 (20.6)	26 (21.3)	
No response	2 (0.7)	2 (1.2)	0 (0.0)	

Mean responses

Overall, mean response scores were similar in both groups for most questions (Figure 2). The mean responses of questions 1, 2, 3 and 9 were significantly lower in the prediction-based group compared to the opinion-based group using a student's t-test and only questions 2 and 9 remained significantly lower when using the MANOVA approach (Supplemental table S2).

Table 2. Proportion of participants giving low (1 and 2) or high (4 and 5) scores in each group

Question	Percentages of low scores (1 and 2) in each group					Percentages of high scores (4 and 5) in each group				
	Opinion-based		Prediction-based	p-value		Opinion-based		Prediction-based	p-value	
	N	%low	N	%low			%high		%high	
1	160	5.00	111	10.71	0.072	160	86.88	111	72.14	0.002
2	160	1.88	108	6.51	0.051	160	97.50	108	80.39	<0.001
3	160	5.00	107	11.67	0.059	160	79.38	107	65.92	0.017
4	160	8.75	107	14.09	0.175	160	71.25	107	65.08	0.313
5	160	8.75	107	11.88	0.367	160	73.75	107	68.81	0.413
6	160	7.50	111	12.83	0.160	160	77.50	111	67.90	0.069
7	160	10.00	102	13.14	0.490	160	73.13	102	65.29	0.199
8	160	6.25	107	10.56	0.231	160	78.13	107	70.50	0.138
9	160	3.75	102	8.17	0.151	160	90.63	102	75.54	0.001
10	160	29.38	109	27.77	0.741	160	38.75	109	41.98	0.571

The distribution of *low* scores (i.e., strongly disagree (score 1) and disagree (score 2) aggregated) did not differ significantly between the opinion-based and the prediction-based groups (Table 2). However, the distribution of *high* scores (i.e., agree (score 4) and strongly agree (score 5) aggregated) was different between both groups for questions 1, 2, 3 and 9. For these questions, the mean percentages of high scores were lower in the prediction-based method. Respondents in the opinion group were more likely to agree or strongly agree with the statement that they liked their method of evaluation compared to the respondents in the prediction group (Likert means 3.89 ± 0.92 and 2.80 ± 1.09, respectively; $p < 0.001$).

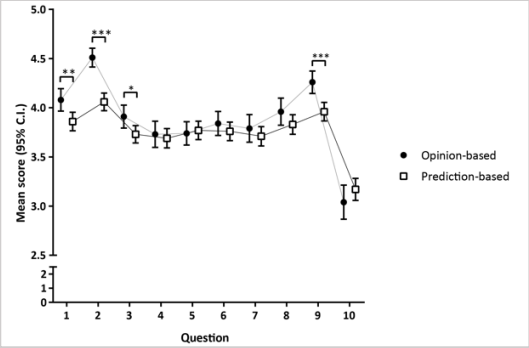

Figure 2. Mean weighted scores per question. [b]

Minimum numbers of required respondents

Overall, the prediction condition required considerably fewer respondents for reliable outcomes than the opinion condition. That is, on average 33±1 respondents were required to come to a sufficiently close approximation of the overall response distribution in the prediction condition, versus an average of 89±7 respondents in the opinion-condition. In the prediction condition, this corresponds to about a third of the actual respondents, versus more than half of the respondents in the opinion condition. For most of the questions, with respect to the numbers of respondents needed for reliable outcomes, the differences between conditions were statistically significant (Table 3).

Table 3. Comparison of the required numbers of respondents in the opinion- and prediction condition

Question	Condition	Number of respondents	Mean number (%) required respondents	Mean difference in required number (95% C.I.)	p-value
1	Opinion	160	85 (53%)	51 (-4; 105)	0.068
	Prediction	111	34 (30%)		
2	Opinion	160	78 (49%)	44 (-12; 103)	0.136
	Prediction	108	34 (32%)		
3	Opinion	160	87 (55%)	55 (2; 103)	**0.031**
	Prediction	107	32 (30%)		
4	Opinion	160	92 (57%)	59 (3; 105)	**0.024**
	Prediction	106	33 (31%)		
5	Opinion	160	86 (54%)	51 (-4; 101)	0.055
	Prediction	107	35 (33%)		
6	Opinion	160	88 (55%)	55 (2; 106)	**0.036**
	Prediction	111	33 (29%)		
7	Opinion	160	93 (58%)	60 (6; 105)	**0.018**
	Prediction	102	33 (33%)		
8	Opinion	160	93 (58%)	59 (5; 111)	**0.030**
	Prediction	107	34 (32%)		
9	Opinion	160	84 (53%)	51 (-5; 106)	0.073
	Prediction	102	33 (33%)		
10	Opinion	160	102 (64%)	70 (20; 113)	**0.003**
	Prediction	109	32 (30%)		

Bootstrap technique with 1000 repetitions to estimate statistical uncertainty around the minimum required respondent numbers. Results were considered stable once the average distribution in the subsample differed less than 5% compared to the average distribution in the whole sample. The number of respondents in the subsample at the point stable results were reached was considered the minimum required number of respondents in this bootstrap sample. C.I., confidence-interval.

Discussion

As expected, based on previous research in undergraduate medical education,[14-17] we found that the prediction-based method required fewer respondents for comparable outcomes in a different educational setting, namely CME targeting a cohort of predominantly primary-care physicians. Moreover, we also found that the mean scores in both conditions were similar for most questions and that

[b] Mean (95% confidence-interval) scores for each question. Prediction-based outcomes have a tendency towards the midpoint of the scale in comparison with the opinion-based outcomes. Mean scores of questions 1, 2, 3, and 9 were significantly higher in the opinion-based group, than the prediction-based group, but answering patterns of respondents in both conditions were similar (compare both lines). Moreover, the questions that received the highest and lowest scores respectively in the opinion-based condition also received the highest and lowest scores in the prediction-based condition. * p-value <0.05, ** p-value <0.01, *** p-values <0.001).

the answering patterns of respondents in both methods were similar: the questions that received the highest and lowest scores respectively in the opinion-based method also received the highest and lowest scores in the prediction-based method. Thus, our findings provide additional empirical evidence for the use of the prediction-based method to obtain valid evaluation outcomes in CME and for Hofstee's line of reasoning that averaged predictions of outcomes lead to better outcomes than averaged opinions.[12,13]

This study adds to previous research on prediction-based course evaluation, by demonstrating its validity and applicability in CME, a context that might complicate the prediction of course evaluation outcomes: CME participants generally come from different work contexts and have different levels of knowledge and professional experience, which may influence what they hope, wish, and expect to get out of the course and how they value the course. Besides, CME participants are not as acquainted with their peers and, therefore, may have less sense of the '*communis opinio'* than undergraduate students do. Despite these characteristics of the CME context, which may complicate rendering adequate predictions, our outcomes support the prediction-based method as an accurate and efficient approach for CME course evaluation.

We discovered several practical issues that should be considered before implementing the prediction-based method. First, prediction-based outcomes tend to gravitate towards the midpoint of the scale in comparison with the opinion-based outcomes. This finding is in line with previous research.[14,15] The majority of respondents estimated that at least a few of their peers would not share the same opinion, spreading the estimated weights across multiple response options, reducing the numbers of extreme scores (i.e. 1, strongly disagree or 5, strongly agree). They did so even if they estimated that most of their peers would strongly agree or strongly disagree. For example, the mean score of question 2 was significantly lower in the prediction-based method, compared with the mean score obtained via the opinion-based method. In the opinion-based group, 98% of respondents gave a Likert score of 4 or 5, resulting in a relatively high mean score of 4.5, indicating that the course was generally well-received. In contrast, in the prediction-based method peers estimated that a significantly lower proportion (80%) of the participants would give a Likert score of 4 or 5, resulting in a significantly lower mean score of 4.1. Despite this difference, in both methods, question 2 received the highest rating. Moreover, the score pattern for all questions was similar in both methods: the same aspects were identified as features of the course that should be retained (higher rated) or that deserved attention (lower rated), even though the scores obtained through the opinion-based method were more extreme. This outcome suggests that applying the prediction-based method may yield valid outcomes but requires a recalibration of how the mean scores should be interpreted. For instance, it could be possible that a lower prediction-based mean (such as 4.1) should be interpreted in the same way (excellent score) as a higher opinion-based mean (such as 4.5), even though the absolute numerical value is lower. Future research might delve deeper into this recalibration process by employing mixed-method approaches, integrating qualitative insights from participants. Additionally, employing statistical methods such as equivalence testing could provide a quantitative basis for recalibration, rigorously assessing the equivalence between prediction-based and opinion-based mean scores and, for instance, examining whether there is more regression to the mean. Moreover, exploring the perceptions and decision-making processes of stakeholders involved in interpreting these scores could enrich our understanding of the nuances involved.Providing participants with the opportunity to provide more explicit feedback, for example by including open questions, may add to the interpretation of the outcomes and the utility of the questionnaire.

Another practical consideration is that the prediction-based method has considerable room for improvement in terms of increasing user-friendliness. Participants appreciated the opinion-based method more than the prediction-based method and various responses had to be excluded for analysis as they did not add up to 100%. This miscalculation problem was also encountered in previous studies.[14,15] Possibly, the appeal made on respondents with regard to estimating the distribution of their peers' perceptions using a new method, as well as calculating correctly, caused too much cognitive load.[14,15] The larger amounts of time and concentration required to answer the more complicated prediction-based format are important challenges that need to be overcome. To improve the user-friendliness of the prediction-method, we suggest using a digital system with automatic calculation to facilitate the prediction task. Such a system could facilitate the prediction task by showing with each question the remaining percentage out of 100% and prevent miscalculations. In addition, the prediction task may be easier to use over time as respondents familiarize themselves with the method and the aspects on which they are invited to evaluate their peers' opinions.

It is unknown which of the two methods yields scores that represent the truth best and which outcomes may best serve as the 'gold standard': the prediction-based method or averaged opinion-based method. Following Hofstee's line of arguing based on Spearman and Brown's axioms that averaged predictions yield better outcomes than averaged opinions,[12,13] the prediction-based outcomes would represent the truth better than the average of personal opinions. The question is whether the prediction assignment succeeds in making students give opinions on behalf of their peers rather than base their predictions on their own experiences. If so, one would expect that respondents, for instance, use the way their peers tend to talk about the course as a frame of reference for their predictions. However, it may also be that respondents use their own course perceptions as a basis for their prediction and simply spread their estimates around these perceptions or predict scores by reflecting their own opinions on different moments during the course. Future studies should determine which method represents the truth best. In-depth qualitative research is needed to unravel the strategies, mechanisms, values, and beliefs behind the prediction process by which participants make their predictions and the frames of reference they use.

Based on the outcomes with the current sample, we estimate that around 30–35 respondents are needed for valid predictions of the 100+ participant course participants' general opinions. This sample size is similar to the number described in the setting of undergraduate medical education.[14,15] It is unknown whether this estimation also holds for larger or smaller samples. In the original election studies of Hofstee and Schaapman, only a fraction of the entire population was needed for reliable outcomes, indicating that there may be a lowerbound with respect to the minimum numbers of respondents needed. Interestingly, 30-35 respondents is similar to the number that Central Limit Theorem suggests.[23] Future research is needed to determine whether the minimum numbers of respondents needed is rather stable with differing sample sizes, to which extent these numbers depend on sample size and whether various methods all require approximately 30 respondents.

A limitation of our study was that the response rates were lower in the prediction group, which could be attributed to the novelty of the method and the difficulty of the prediction task. In addition, despite random allocation, the respondent characteristics were different between both groups. One might therefore question whether the inferences from the traditional group are comparable with those of the prediction group, as the participants, consequently, may have had differing opinions on the items. Although we cannot exclude that the lower response rates and different respondent characteristics may have resulted in biased outcomes, there are several indications that sustain the validity and use of the prediction outcomes. First, both methods resulted in similar outcomes and even with clearly smaller response rates, the prediction-based method identified the same course aspects as most and least positive as the opinion-based method, despite differences between the participant samples in both conditions. What strengthens the validity of our outcomes additionally, is that the mechanism of action behind the prediction-based approach–that prediction reduces bias due to personal factors such as prejudices, emotions, and personally irrelevant thoughts–not only evident in the original election context but also in undergraduate education,[12-15] was also found in our postgraduate education context despite its distinct characteristics including a heterogeneous participant group, limited peer interaction, and a lack of '*communis opinion.*' Taken together, these data suggest that including respondents from subsamples with different characteristics when applying the prediction-based approach will not lead to biased outcomes, and that the prediction-based evaluation can be organized by randomly selecting respondents from the population.[15] Moreover, we conclude that the prediction method is also a useful approach to obtain valid evaluation outcomes in the CME context. Important strengths of this study were that groups were randomized, and that the prediction-based method was studied in a novel educational setting, namely a heterogeneous group of physicians participating in CME. Another strength is that we used more advanced statistical methodology than previous studies, which strengthens the scientific support that we found for the use of the prediction-based method for evaluation of medical courses, and CME in particular.

Conclusion

We found further evidence for the validity of the prediction-based evaluation method in the context of CME, an educational setting that might complicate the prediction of evaluation outcomes as it involved physicians from across the country differing in age, working contexts and levels of experience. Furthermore, by using more advanced statistical methodology than previous studies on this issue did, we strengthened the empirical evidence for the use of the prediction-based method for evaluation purposes in medical education. Our study demonstrates that also in a heterogeneous setting like CME, the prediction-based

method required fewer respondents for comparable outcomes regarding the strengths of the course and the aspects that might need improvement. The application of the prediction-based method showed room for improvement in terms of user-friendliness and requires a recalibration regarding the interpretation of the numerical score, due to a tendency towards the midpoint of the scale, possibly reflecting the spread of estimations over multiple response options due to uncertainty on the opinion of the entire group.

Conflicts of Interest: The authors do not have competing interests.
Funding: None
Authorship: JC, MvD, and MT made a substantial contribution to the design of the study, analysis, and interpretation of the data. JC worked on initial drafts of the manuscript. JSA contributed to the methodological design of the study and several drafts of the manuscript. FD supervised the statistical analysis. JB supervised the study, contributing to the design of the study, the interpretation of the data, and modification of the manuscript. All authors were involved in finalizing the draft and approved the final manuscript for publication. All authors confirm being accountable for all aspects of the work and for the accuracy and integrity of the content.
Acknowledgements: We thank the Educational Board of the Boerhaave Committee for supporting the initiative to conduct this study, Marjan Hornstra-Moedt and Tamara van Battum-Klink for their enormous practical support in gathering the data for this study, and Nikki Kromkamp for her valuable contribution to this manuscript.
Edited by: Lawrence Grierson (section editor); Christina St. Onge (senior section editor); Marcel D'Eon (editor-in-chief)

References

1. Etter JF, Perneger TV. Analysis of non-response bias in a mailed health survey. *J Clin Epidemiol* 1997;50:1123-1128. https://doi.org/10.1016/S0895-4356(97)00166-2
2. Asch DA, Jedrziewski MK, Christakis NA. Response rates to mail surveys published in medical journals. *J Clin Epidemiol* 1997;50:1129-1136. https://doi.org/10.1016/S0895-4356(97)00126-1
3. Johnson T, Owens L. *Survey response rate reporting in the professional literature.* Proceedings of the 58th Annual Meeting of the American Association for Public Opinion Research, Nashville; 2003.
4. Yu J, Cooper H. A quantitative review of research design effects on response rates to questionnaires. *J Market Res.* 1983:36-44. https://doi.org/10.1177/002224378302000105
5. Spooren P, Brockx B, Mortelmans D. On the validity of student evaluation of teaching: The state of the art. *Rev Educ Res.* 2013;83:598-642. https://doi.org/10.3102/0034654313496870
6. Rodriguez E, Marquett R, Hinton L, McBride M, Gallagher-Thompson D. The impact of education on care practices: an exploratory study of the influence of "action plans" on the behavior of health professionals. *Intern psychogeriat.* 2010;22:897-908. https://doi.org/10.1017/S1041610210001031
7. Skubleny D, Switzer N, Karmali S, de Gara C. Endoscopy services and training: a national survey of general surgeons. *Can J Surg.* 2015;58:330-334. https://doi.org/10.1503/cjs.015914
8. Van Geest JB, Johnson TP, Welch VL. Methodologies for improving response rates in surveys of physicians: a systematic review. *Eval Health Profess.* 2007;30:303-321. https://doi.org/10.1177/0163278707307899
9. Yarger JB, James TA, Ashikaga T, et al. Characteristics in response rates for surveys administered to surgery residents. *Surgery* 2013;154:38-45. https://doi.org/10.1016/j.surg.2013.04.060
10. Asch S, Connor SE, Hamilton EG, Fox SA. Problems in recruiting community-based physicians for health services research. *J Gen Intern Med.* 2000;15:591-599. https://doi.org/10.1046/j.1525-1497.2000.02329.x
11. Brouns JW, Berkenbosch L, Ploemen-Suijker FD, Heyligers I, Busari JO. Medical residents perceptions of the need for management education in the postgraduate curriculum: a preliminary study. *Int J Med Educ* 2010;1:76-82. https://doi.org/10.5116/ijme.4cd8.43f1
12. Hofstee W. Uitbuiting van de dagelijkse ervaring: Facetten van een psychometrische waarheidsbenadering [Exploitation of the daily experience]. *Nederlands Tijdschrift voor de Psychologic* 1993;48:277-287.
13. Hofstee W, Schaapman H. Bets beat polls: Averaged predictions of election outcomes. *Acta Politica* 1990;25:194-207.
14. Cohen-Schotanus J, Schönrock-Adema J, Schmidt HG. Quality of courses evaluated by 'predictions' rather than opinions: Fewer respondents needed for similar results. *Med Teach.* 2010;32:851-856. https://doi.org/10.3109/01421591003697465
15. Schönrock-Adema J, Lubarsky S, Chalk C, Steinert Y, Cohen-Schotanus J. 'What would my classmates say?'An international study of the prediction-based method of course evaluation. *Med Ed.* 2013;47:453-462. https://doi.org/10.1111/medu.12126
16. Tomes T, Coetzee S, Schmulian A. Prediction-based student evaluations of teaching as an alternative to traditional opinion-based evaluations. *Assess Eval Higher Educ.* 2019:1-15. https://doi.org/10.1080/02602938.2019.1594157
17. Bacchi S, Guo B, Brown S, Symonds I, Hudson JN. Can Australian medical students' predictions of peers' responses assist with gaining reliable results on course evaluations? *Focus Health Profess Educ.* 2018;19:14. https://doi.org/10.11157/fohpe.v19i2.250
18. Abrahams MB, Friedman CP. Preclinical course-evaluation methods at U.S. and Canadian medical schools. *Acad Med* 1996;71:371-374. https://doi.org/10.1097/00001888-199604000-00015.
19. Griffin P, Coates H, Mcinnis C, James R. The development of an extended course experience questionnaire. *Qual Higher Educ.* 2003;9:259-266. https://doi.org/10.1080/135383203200015111
20. Norman G. Likert scales, levels of measurement and the "laws" of statistics. *Adv Health Sci Educ Theory Pract* 2010;15:625-632. https://doi.org/10.1007/s10459-010-9222-y
21. le Cessie S, Goeman JJ, Dekkers OM. Who is afraid of non-normal data? Choosing between parametric and non-parametric tests. *Euro J Endocrinol.* 2020;182:E1-E3. https://doi.org/10.1530/EJE-19-0922
22. Altman DG, Bland JM. How to obtain the P value from a confidence interval. *Bmj*. 2011;343:d2304. https://doi.org/10.1136/bmj.d2090
23. Louangrath P. Minimum sample size method based on survey scales. *Int J Res Methodol Soc Sci* 2017;3:44-52.