Universal Journal of Educational Research 1(3): 247-254, 2013 http://www.hrpub.org
DOI: 10.13189/ujer.2013.010315

Surgical Theatre (Operating Room) Measure STEEM (OREEM) Scoring Overestimates Educational Environment: the 1-to-L Bias

Ioannis DK Dimoliatis[1], Eleni Jelastopulu[2,*]

[1]Department of Hygiene and Epidemiology, University of Ioannina Medical School, Ioannina, Greece
[2]Department of Public Health, School of Medicine, University of Patras, Patras, Greece
*Corresponding Author: jelasto@upatras.gr

Abstract The surgical theatre educational environment measures STEEM, OREEM and mini-STEEM for students (student-STEEM) comprise an up to now disregarded systematic overestimation (OE) due to inaccurate percentage calculation. The aim of the present study was to investigate the magnitude of and suggest a correction for this systematic bias. After an initial theoretical exploration of the problem, published scores were retrieved from the literature and corrected using statistical theorems. Overestimations and differences between pseudo-percentages and real percentages were plotted against real percentages. Reported STEEM overall mean score of 74.4% (pseudopercentage) was corrected to 67.9% (real percentage), eliminating thus the 6.4% OE. Corresponding figures for OREEM and student-STEEM are 73.6%, 67.0%, 6.6% and 69.1%, 61.4%, 7.7% respectively. A total of 45 overestimated scores were retrieved and corrected. OE (range 2.8 to 13.6%, mean 7.3%) showed a complete ($r = -1$, $p < 0.001$) negative linear regression of real percentages (RP), namely, OE = 20-0.2*RP. No uncorrected score can achieve less than 20%. The non-0-based 1-to-5 coding overestimates STEEM, OREEM and student-STEEM educational environment scores if expressed as percentages due to the '1-to-5 bias', or rather 1-to-L bias, whereupon L correlates to the number of points in the Likert scale, the number of options. The worse the educational environment the greater the overestimation, reducing instruments' usefulness exactly then when alarm bells should be ringing. Hence, question coding should always be zero (0) based, as proposed by Likert. The 1-to-L bias applies to any questionnaire at any field of research.

Keywords Educational Environment, Likert Scale, STEEM / OREEM Questionnaire, Overestimation, Bias, Scoring

1. Introduction- Theoretical Exploration

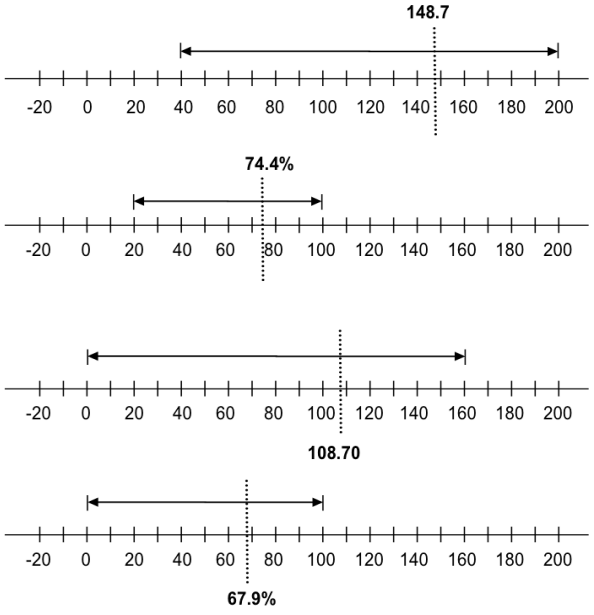
Many questionnaires assessing educational environment, as perceived by the participants, have been developed. The DREEM for undergraduates [1, 2], the PHEEM for hospital-based junior doctors [3], the ATEEM for anesthetists in the surgical theatre [4], the STEEM [5] and OREEM [6] for surgeons in the surgical theatre / operating room, and the mini-STEEM [7], a short version of STEEM for undergraduates, hereinafter referred to as student-STEEM or sSTEEM. DREEM, PHEEM and ATEEM use a five-point 0-to-4 Likert scale to code individual questions. On the contrary, the other three (STEEM, OREEM and sSTEEM) use the five-point 1-to-5 scale. However, this raises a problem when the scores are expressed as percentages. Namely, it introduces error into the assessment by overestimating the quality of the educational environment, especially when it is (very) poor.

DREEM consists of 50 questions, each scored 0-to-4 (strongly disagree to strongly agree), thus giving an overall score range 0-to-200 (50*0 to 50*4). PHEEM and ATEEM consist of 40 questions, each scored 0–4, giving an overall score range 0–160 (40*0–40*4).[1] STEEM and OREEM consist of 40 questions too, but they are scored 1–5, giving an overall score range 40–200 (40*1–40*5), and sSTEEM consists of 13 questions, each scored 1–5, giving an overall score range 13–65 (13*1–13*5). Each of the inventories is divided in a different number of subscales containing a different number of questions, giving a lot of subscale score ranges. To interpret a score obtained after administrating any of the instruments, the score range is usually divided in four equal zones, the lower of which representing the very poor educational environment, the second the poor, the third the good, and the fourth the very good [2]. But one has to remember all these score ranges, interpretation zones and cut-points.

To prevent confusion, it is a usual practice to transform the actual ranges into the standard 0–100 scale ('the standard scoring method' [8, 9]) and interpret any individual score as

[1] Hereinafter * denotes the sign of multiplication, and '–' the '-to-'.

very poor if it lies within the 0–24.9 zone, poor in 25–49.9, good in 50–74.9, and very good in 75–100. However, this transformation has some pitfalls if individual questions are scored 1–5, as it happens in STEEM, OREEM and sSTEEM, which might distort participants' perceptions. Original papers report overall mean score (OMS) "148.7/200 (74.4%)" [5], "147.2/200 (73.6%)" [6], and "44.9/65 (69.1%)" [7]. Obviously, these 'percentages' are the quotients of the corresponding divisions: 148.7/200 = 0.7435, 147.2/200 = 0.736, and 44.9/65 = 0.6908. That is, the actual OMSs were divided by the upper limit of the corresponding range, overlooking that its lower limit was not zero, but 40 in STEEM and OREEM and 13 in sSTEEM. However, a true percentage equals the OMS divided by the upper limit, if and only if the lower limit is zero. Otherwise, the quotient is a pseudo-percentage, not a percentage. The expression "148.7/200 (74.4%)" is quite misleading; more accurately, it should be "148.7 in the range 40–200 (or 74.4 in the range 20–100)", but not 74.4%. Anybody seeing "148.7" or any other number automatically understands a point within a range from 0 to an upper limit and anybody seeing a percentage automatically understands a point within the standard range 0–100%. However, 74.4% is a point within the 20–100 range, i.e., it is not a percentage really, and this causes the problem (Figure 1).



Double arrow: Score range on the real number line.
Dotted vertical line: Overall mean score (OMS) on the real number line.
1st scale: the 1-based non-standard range 40–200 and the 1-based non-standard OMS 148.7 (non-0-anchored score.)
2nd scale: the 1-based pseudostandard range 20–100 and the 1-based pseudostandard OMS 74.4% (non-0-anchored pseudopercentage).
3rd scale: the 0-based non-standard range 0–160 and the 0-based non-standard OMS 108.7 (0-anchored score).
4th scale: the 0-based standard range 0–100 and the 0-based standard OMS 67.9%. (0-anchored real percentage).
1st and 2nd scale: Reported scores and percentages (what is reported).
3rd and 4th scale: Corrected scores and percentages (what should be).

Figure 1. Pseudoscores and Pseudopercentages versus Real Scores and Real Percentages

Using the reported STEEM OMS as example, the first two graphs in Figure 1 clarify what has been reported, while the next two what should be reported.

In the top graph, the arrow extending from 40 to 200 indicates the non-standard 40-based overall STEEM range, while the dotted vertical line at 148.7 indicates the non-standard 40-based OMS. In the second graph, their transformation to supposed standard percent values is presented; the arrow extending from 20 to 100 (100*40/200 to 100*200/200) indicates the pseudostandard 20-based overall STEEM range, and the dotted vertical line at 74.4 indicates the reported as standard but 20-based pseudostandard OMS. In fact, the worst fifth (0–20) of the real standard range 0–100 has been cut and the reported as standard values start from 20%, i.e., they have been shrunk to the right and thus they are pseudo-standard. That's where overestimation comes from. In the third graph, the arrow and the dotted vertical line have been moved to the left by exactly 40 points. The arrow now indicates the non-standard 0-based overall STEEM range 0–160, while the dotted vertical line at 108.7 indicates the non-standard 0-based OMS. All these new points equal the reported ones minus 40 (0 = 40-40; 160 = 200-40; 108.7 = 148.7-40). In the last graph, the 0–160 range has been shrunk to fit the standard 0-based 0–100 range, where the OMS becomes 67.9 (108.7*100/160 = 67.9375), i.e., the 0-based standard OMS (67.9) equals the 0-based actual OMS (108.7) multiplied by the constant C = 100/160. Therefore, the reported non 0-based pseudostandard OMS 74.4% overestimates the 0-based standard OMS 67.9% by 6.4% (= 74.35% - 67.9375%).

Table 1. Five different scenarios from the worst (A) to the best (E) perceived education environment

SCENARIO [1]		A	B	C	D	E
AO	Strongly Disagree	40	0	0	0	0
	Disagree	0	40	0	0	0
	Uncertain	0	0	40	0	0
	Agree	0	0	0	40	0
	Strongly Agree	0	0	0	0	40
1-5 coding	40-based initial OMS[2]	40	80	120	160	200
	Pseudostand OMS[3]	20	40	60	80	100
0-4 coding	0-based initial OMS[4]	0	40	80	120	160
	Real standard OMS[5]	0	25	50	75	100
OE	Overestimation[6]	20	15	10	5	0
	Reduction rate[7]	-0.2	-0.2	-0.2	-0.2	

AO: Answer option distribution; OE: Overestimation
[1] A participant or a set of participants choose the same option in all forty questions, either exclusively 'strongly disagree' (scenario A) or exclusively 'disagree' (B) or exclusively 'uncertain' (C) or exclusively 'agree' (D) or exclusively 'strongly agree' (E).
[2] 40*L, L = 1, 2, 3, 4, 5 for scenarios A, B, C, D, E respectively.
[3] 100*(40-based OMS)/200.
[4] 40*L, L = 0, 1, 2, 3, 4 for scenarios A, B, C, D, E respectively.
[5] 100*(0-based OMS)/160.
[6] (Pseudostandard OMS) minus (Real standard OMS).
[7] (15-20)/(25-0) = (10-15)/(50-25) = (5-10)/(75-50) = (0-5)/(100-75) = -5/25 = -0.2 = constant (deliberately put between scenarios), the b coefficient in formula {1}.

The problem originated when the 0-based 0–4 coding had been moved to the right by just 1 point (1 instead of 0, 2 instead of 1, etc.) and the 1–5 range was obtained. Then, adding 40 questions to produce the 40-question overall score, the 0–160 range was moved by 40 points and the 40–200 range was obtained. The consequences are explored in Table 1, using five different scenarios in a STEEM administration to a single participant (or a group of participants) who selects exclusively the same option in all forty questions, either 'strongly disagree' (scenario A) or 'disagree' (B) or 'uncertain' (C) or 'agree' (D) or 'strongly agree' (E).

Coding the answers 1–5 as the papers in question did, the 40-based actual OMS is 40 (scenario A), 80 (B), 120 (C), 160 (D) and 200 (E). Dividing them by the maximum score possible (200), the supposed standard but in reality pseudostandard OMS becomes 20, 40, 60, 80 and 100 respectively (Table 1). Obviously, no such OMS can be less than 20%, the score for the worst imaginable environment where the participant chose 'strongly disagree' in all questions (scenario A). The score range is not 0%–100%, but only 20%–100%: the worst fifth has been cut, i.e., any non-0-based 'standard' OMS is a pseudostandard.

Coding the answers 0–4, as DREEM, PHEEM and ATEEM did (and STEEM, OREEM and sSTEEM should), the initial non-standard OMS becomes 0, 40, 80, 120 and 160 respectively, and the corresponding (real) standard OMS 0%, 25%, 50%, 75% and 100%, ranging in a real percent scale. The differences between pseudostandard and standard OMS (namely, 20%, 15%, 10%, 5%, 0% respectively) are the corresponding overestimations, decreasing from 20% to 0% as the standard OMS increases from 0% to 100%, with a constant rate of -0.2, the minus sign indicating their reverse relation: the greater the standard OMS the lesser the overestimation and vice-versa. The maximum overestimation can be 20% and the minimum 0% in the worst and the best imaginable environment respectively (scenarios A and E, with standard OMS 0 and 100). In other words:

$$\text{Overestimation} = (20\%)-(0.2)*(\text{standard OMS}) \quad \{1\}$$

Having revealed and explained the overestimation introduced by the 1-to-5 bias, our aim was to correct all reported scores in the original papers and explore the degree of over-appraisal.

2. Materials and Methods

2.1. Correction

All overall, subscale and question pseudostandard scores anywhere in the original papers [5-7,10] were retrieved (Table 2) and corrected, using the following statistical theorems [11]:

$$M(C+X) = C+M(X) \quad \{2\}$$

$$SD(C+X) = SD(X) \quad \{3\}$$

$$M(CX) = CM(X) \quad \{4\}$$

$$SD(CX) = |C|SD(X) \quad \{5\}$$

That is, if a constant C is added to all individual values of a variable X, the mean (M) of the new variable C+X equals the mean of variable X plus C {2}, while the standard deviation (SD) of C+X equals the SD of X {3}. And if all individual values of a variable X are multiplied by a constant C, the mean of variable CX equals the mean of variable X multiplied by C {4}, while the SD of CX equals the SD of X multiplied by the absolute value of C {5}.

The first two formulas were used to rescale the non-0-based (40-based, 13-based, 1-based etc) reported mean scores and standard deviations to 0-based values, where C = -Q (Q the number of questions per scale or subscale): subtracting Q from the reported non-0-based actual scores, the 0-based actual scores were obtained due to theorem {2}, while the 0-based standard deviations equal the reported non-0-based standard deviations {3}. The next two theorems were used to transform these 0-based values to the standard range 0–100, where C = 25/M (see notes in Table 2 for details): multiplying the 0-based values by 25/M their equivalents in the standard range 0–100 were obtained due to {4} and {5}. Finally, since it should remain unchanged in reported and corrected data, the coefficient of variance (CV = SD/M) was used to verify our transformations. The graphical presentation of these theorems is demonstrated in Figure 1.

2.2. Overestimation and Interpretation

The overestimation was calculated as the difference between reported non-0-based pseudostandard mean scores and corrected 0-based standard mean scores, and regressed against corrected standard mean scores. Dividing the standard scale 0–100 in four equal zones, 0–24.9, 25–49.9, 50–74.9, 75–100 [2], we compared the distribution in these zones of the pseudostandard and real standard mean scores. The same distribution of 33 DREEM standard overall mean scores from a recent review [12] was also compared to both.

Table 2. All reported non-0-based scores, corrected 0-based scores, and calculated overestimations

Scale/ Subscale/ Question	L	Q	Reported non-0-based Scores						*Corrected 0-based Scores*						*OE*
			B_n	U_n	M_n	SD_n	$M_{n\%}$	$SD_{n\%}$	B_0	U_0	M_0	SD_0	$M_{0\%}$	$SD_{0\%}$	
OREEM (Kanashiro et al, 2006)															
Overall	5	40	40	200	147.2		73.6		*0*	*160*	*107.2*		*67.0*		*6.6*
Males	5	40	40	200	150.7		75.4		*0*	*160*	*110.7*		*69.2*		*6.2*
Females	5	40	40	200	136.8		68.4		*0*	*160*	*96.8*		*60.5*		*7.9*
Hospital PLC	5	40	40	200	154.8		77.4		*0*	*160*	*114.8*		*71.8*		*5.7*
Hospital FMC	5	40	40	200	142.7		71.4		*0*	*160*	*102.7*		*64.2*		*7.2*
S1 Teaching & training	5	13	13	65	47.5		73.1		*0*	*52*	*34.5*		*66.3*		*6.7*
S2 Learning opportunities	5	11	11	55	39.9		72.5		*0*	*44*	*28.9*		*65.7*		*6.9*
S3 Atmosphere	5	8	8	40	31.1		77.9		*0*	*32*	*23.1*		*72.3*		*5.5*
S4 Workload/Supervision/Support	5	8	8	40	27.0		67.5		*0*	*32*	*19.0*		*59.4*		*8.1*
S4a Workload/Super/Sup Juniors	5	8	8	40	25.5		63.8		*0*	*32*	*17.5*		*54.7*		*9.1*
S4b Workload/Super/Sup Seniors	5	8	8	40	29.2		73.0		*0*	*32*	*21.2*		*66.3*		*6.8*
Qi [scores not given]	5	1	1	5					*0*	*4*					
STEEM (Cassar, 2004)															
Overall	5	40	40	200	148.7		74.4		*0*	*160*	*108.7*		*67.9*		*6.4*
S1 Teaching & training	5	13	13	65	*51.3*		78.9		*0*	*52*	*38.3*		*73.6*		*5.3*
S2 Learning opportunities	5	11	11	55	*37.1*		67.5		*0*	*44*	*26.1*		*59.4*		*8.1*
S3 Atmosphere	5	8	8	40	*30.4*		76.0		*0*	*32*	*22.4*		*70.0*		*6.0*
S4 Workload/Supervision/Support	5	8	8	40	*30.0*		75.0		*0*	*32*	*22.0*		*68.8*		*6.3*
Q2 I get on well with my trainer	5	1	1	5	4.4		*88.8*		*0*	*4*	*3.4*		*86.0*		*2.8*
Q6 Trainer's surgical skills are good	5	1	1	5	4.4		*88.8*		*0*	*4*	*3.4*		*86.0*		*2.8*
Q39 Supervision adequate my level	5	1	1	5	4.4		*88.0*		*0*	*4*	*3.4*		*85.0*		*3.0*
Q20 Sufficient emergency procedures	5	1	1	5	2.8		*56.0*		*0*	*4*	*1.8*		*45.0*		*11.0*
Q27 Nurses dislike when I operate	5	1	1	5	2.5		*49.6*		*0*	*4*	*1.5*		*37.0*		*12.6*
Q38 I get bleeped during operations	5	1	1	5	2.3		*45.6*		*0*	*4*	*1.3*		*32.0*		*13.6*
Overall (Nagraj et al, 2006)	5	40	40	200	139		*69.5*		*0*	*160*	*99.0*		*61.9*		*7.6*
Student-STEEM (Nagraj et al, 2006)															
Overall	5	13	13	65	44.9	7.1	*69.1*	*10.9*	*0*	*52*	*31.9*	*7.1*	*61.4*	*13.6*	*7.7*
S1 Good surgical operating experience	5	5	*5*	*25*	*14.7*		*58.9*		*0*	*20*	*9.7*		*48.6*		*10.3*
S2 Friendly atmosphere in theatre	5	4	*4*	*20*	*15.3*		*76.5*		*0*	*16*	*11.3*		*70.6*		*5.9*
S3 Discrimination against me	5	3	*3*	*15*	*12.0*		*80.0*		*0*	*12*	*9.0*		*75.0*		*5.0*
Q1 Enthusiastic trainer	5	1	1	5	3.7	1.0	*74.5*	*20.7*	*0*	*4*	*2.7*	*1.0*	*68.1*	*25.8*	*6.4*
Q2 Theatre staff friendly	5	1	1	5	3.9	0.8	*78.6*	*16.2*	*0*	*4*	*2.9*	*0.8*	*73.2*	*20.3*	*5.4*
Q3 Enough theatre sessions	5	1	1	5	3.9	0.9	*78.4*	*18.5*	*0*	*4*	*2.9*	*0.9*	*73.0*	*23.2*	*5.4*
Q4 Trainer discusses techniques	5	1	1	5	2.8	1.1	*55.5*	*22.0*	*0*	*4*	*1.8*	*1.1*	*44.4*	*27.5*	*11.1*
Q5 Right case mix list	5	1	1	5	3.3	1.0	*66.5*	*19.1*	*0*	*4*	*2.3*	*1.0*	*58.1*	*23.9*	*8.4*
Q6 Good emergency cases variety	5	1	1	5	3.0	1.0	*60.8*	*20.8*	*0*	*4*	*2.0*	*1.0*	*51.0*	*26.0*	*9.8*
Q7 Enough opportunity to assist	5	1	1	5	3.1	1.2	*61.0*	*23.8*	*0*	*4*	*2.1*	*1.2*	*51.3*	*29.7*	*9.7*
Q8 Operations too complex for me	5	1	1	5	3.1	1.1	*61.0*	*22.6*	*0*	*4*	*2.1*	*1.1*	*51.3*	*28.2*	*9.7*
Q9 Anaesthetists pressure trainers	5	1	1	5	3.6	0.9	*72.5*	*18.4*	*0*	*4*	*2.6*	*0.9*	*65.7*	*23.0*	*6.9*
Q10 Sex discrimination in theatre	5	1	1	5	4.1	1.0	*81.8*	*19.0*	*0*	*4*	*3.1*	*0.9*	*77.3*	*23.7*	*4.5*
Q11 Race discrimination in theatre	5	1	1	5	4.3	0.9	*85.7*	*18.1*	*0*	*4*	*3.3*	*0.9*	*82.1*	*22.6*	*3.6*
Q12 Too busy doing other work	5	1	1	5	2.5	1.1	*50.6*	*22.7*	*0*	*4*	*1.5*	*1.1*	*38.3*	*28.4*	*12.3*

Q14 Pleasant theatre atmosphere	5	1	1	5	3.7	0.8	*74.5*	*15.7*	*0*	*4*	*2.7*	*0.8*	*68.1*	*19.6*	*6.4*
STEEM (Mahoney et al, 2010)															
Overall	5	40	40	200	*147.6*		*73.8*		*0*	*160*	*107.6*		*67.3*		*6.55*
S1 Teaching & training	5	13	13	65	*46.81*		72.0		*0*	*52*	*33.81*		*65.0*		*7.0*
S2 Learning opportunities	5	11	11	55	*39.57*		72.0		*0*	*44*	*28.57*		*65.0*		*7.0*
S3 Atmosphere	5	8	8	40	*31.2*		78.0		*0*	*32*	*23.2*		*72.5*		*5.5*
S4 Workload / Supervision / Support	5	8	8	40	*30.0*		75.0		*0*	*32*	*22.0*		*68.8*		*6.25*

Abbreviations. In the first column: Si / Qi = the subscale / question i, i = 1, 2, 3, ... In the last column: OE = overestimation. In the paper: OMS / SMS / QMS = overall / subscale / question mean score.

Interpretation. 75-100 very good, 50-74.9 good, 25-49.9 poor, 0-24.9 very poor (no such score was reported).

Symbols: L (in honor of Likert) = the number of points (anchors) of an L-point Likert scale for question coding; in all educational environment measures L=5: 'strongly disagree', 'disagree', 'uncertain', 'agree', 'strongly agree'. Q = the number of questions per scale, subscale or question. B = the bottom (lower) limit of a score range. U = the upper limit of a score range. M = mean score (in bold scores that changed interpretation zone after correction), SD = standard deviation. C = constant; |C| the absolute value of C. Any symbol with a subscript (e.g. M_n, M_0, $M_{n\%}$ etc) denotes the symbol in a non-0-based ($_n$) and 0-based ($_0$), and the standard 0–100 ($_{\%}$) scale.

Calculations: Columns L to $M_{n\%}$ appear as given in the corresponding papers, unless in italics denoting numbers calculated by us. $B_n = Q*1 = Q$. $U_n = QL$. $SD_{n\%} = |100/(QL)|SD_n$, after formula {5}. $B_0 = B_n-Q = Q-Q = 0$. $U_0 = U_n-Q = Q(L-1)$. $M_0 = M_n-Q$, after formula {2}. $SD_0 = SD_n$, after formula {3}. $M_{0\%} = (100/(Q(L-1)))M_0$, after formula {4}. $SD_{0\%} = |100/(Q(L-1))|SD_0$, after formula {5}. $OE = M_{n\%}-M_{0\%}$. All Mahoney et al reported values in italics have been calculated by us on the basis of reported $M_{n\%}$ and the number of items, using formulas (13*72+11*72+8*78+8*75)/40 for the overall $M_{n\%}$ and $(M_{n\%}*U_n)/100$ for each M_n. Although the numbers are shown with one decimal point, all calculations were carried out using the most accurate value provided anywhere in the original papers.

Generalizing, any non-0-based non-standard score containing Q questions, Q = 1, 2, 3, ..., coded 1–L, and therefore ranging Q–QL, before its transformation to a percentage (standard), must first be rescaled to its 0-based equivalent ranging 0–Q(L-1), i.e., to $M_{0–Q(L-1)} = M_{Q–QL}-Q$ and $SD_{0–Q(L-1)} = SD_{Q–QL}$, due to theorems {2} and {3}, and then be standardized to a genuine percentage ranging 0–100, i.e., to $M_{0–100} = (100/(Q(L-1)))M_{0–Q(L-1)}$ and $SD_{0–100} = (100/(Q(L-1)))SD_{0–Q(L-1)}$, as per theorems {4} and {5}. Otherwise it will be a pseudopercentage (pseudostandard) ranging $100((Q–QL)/QL) = 100(1/L–1) = 100/L–100$; i.e., 50-100 if L=2, 20-100 if L=5, 10-100 if L=10 etc.

Verification: Coefficients of variation (CV = SD/M) were as expected in all the cases where SD_n was given: $SD_{n\%}/M_{n\%} = SD_n/M_n$ and $SD_{0\%}/M_{0\%} = SD_0/M_0 = SD_n/(M_n-Q)$; for the simplicity of the Table these coefficients are not shown.

3. Results

3.1. Reported and Corrected Scores

The results are shown in Table 2. Column L presents the number of points in the L-point Likert scale; in all three questionnaires L=5. Column Q presents the number of questions per scale, subscale or question; the OREEM and STEEM consist of 40 questions, their first subscale consists of 13 questions, etc. The next two columns present the lower (B_n) and upper (U_n) limits of the range within which any non-0-based score could be reported; for example, 40–200 for the overall STEEM / OREEM scores, 13–65 for the overall sSTEEM score, 1–5 for any single question score, etc. The columns M_n and SD_n present the reported non-0-based overall, subscale or question non-standard mean scores and standard deviations; for example, the non-standard OMS was 147.2 for OREEM (no SD was reported), 148.7 for STEEM (no SD was reported), and 44.9 for sSTEEM with a SD of 7.1. The columns $M_{n\%}$ and $SD_{n\%}$ present corresponding reported as standard mean scores and standard deviations in the erroneously supposed standard (1–100) but in fact pseudostandard (20–100) scale; for example reported as 'standard' overall means were 73.6%, 74.4% and 69.1%; no SD was reported, except for sSTEEM (10.9% overall, 20.7% first question etc.). The following six columns present the 0-based corrected values. Columns B_0 and U_0 present the lower and upper limits of the range within which any 0-based score could be found. Columns M_0 and SD_0 present the 0-based mean scores and standard deviations, and columns $M_{0\%}$ and $SD_{0\%}$ present mean scores and standard deviations in the standard 0–100 scale. Finally, the last column presents the overestimations (OE).

Here are three examples from Table 2, an overall, a subscale, and a question score. The OREEM reported OMS in the 40-based 40–200 range was 147.2, falsely reported as 73.6%; transformed to the 0-based 0–160 range, they become 107.2 and 67.0% (correct), which therefore indicates an overestimate of 6.6%. The STEEM 'teaching & training' subscale reported mean score in the 13-based 13–65 range was 51.3, falsely reported as 78.9%; transformed to the 0-based 0–52 range, they become 38.3 and 73.6% (correct), a 5.3% overestimation. Finally, the sSTEEM question 12 reported mean score in the 1-based 1–5 range was 2.53 (SD = 1.1), falsely reported as 50.6% (SD = 22.7); corrected to the 0-based 0–4 range, they become 1.53 (SD = 1.1) and 38.3% (SD = 28.4), a 12.3% overestimation.

3.2. Overestimation and Its Implication on Score Interpretation

Figure 2 reveals a complete (r = -1) negative linear relation between the overestimation and the corrected standard mean score ($M_{0\%}$), which is not obvious in the last column (OE) of Table 2. Because of this perfect linearity, we can predict no overestimation at all if $M_{0\%} = 100$ (this makes sense: there is no room for improvement), but there is a 20% overestimation if $M_{0\%} = 0$, i.e., the 1-to-5 bias adds up to 20% overestimation as $M_{0\%}$ moves from 100% to 0%, with a rate of 0.2 per $M_{0\%}$ unit. This is exactly what was theoretically predicted in Table 1 and formula {1}.

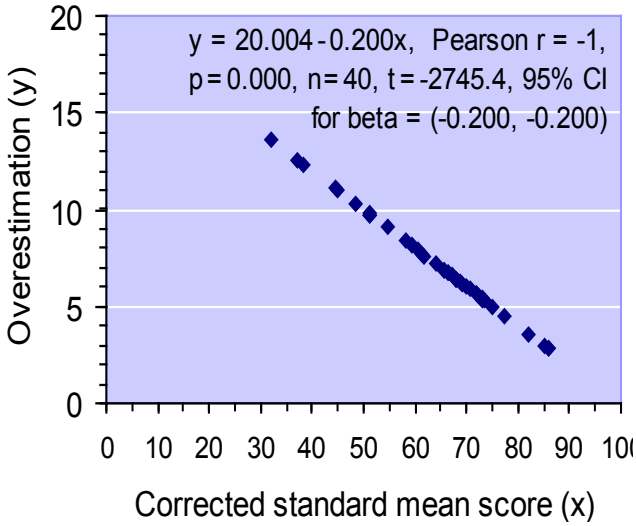


Figure 2. Overestimation against corrected (real) standard mean score ($M_{0\%}$) (Data from Table 2)

Thus, the uncorrected OREEM, STEEM, and sSTEEM scores will never fall within the worst fifth 0–20% (Figure 3), although this might be the situation as appraised by the survey participants. The uncorrected 'standard' mean score, i.e., the pseudostandard mean score ($M_{n\%}$), can never be less than 20%, since this 20% is entirely attributable to the overestimation. In real life, it would almost be impossible to find a score indicating a very poor environment, i.e., within the worst quarter: in order to obtain an uncorrected (pseudostandard) 25% you would only need a real standard score of 6%, the 19% overestimation makes up the rest (Figure 3).

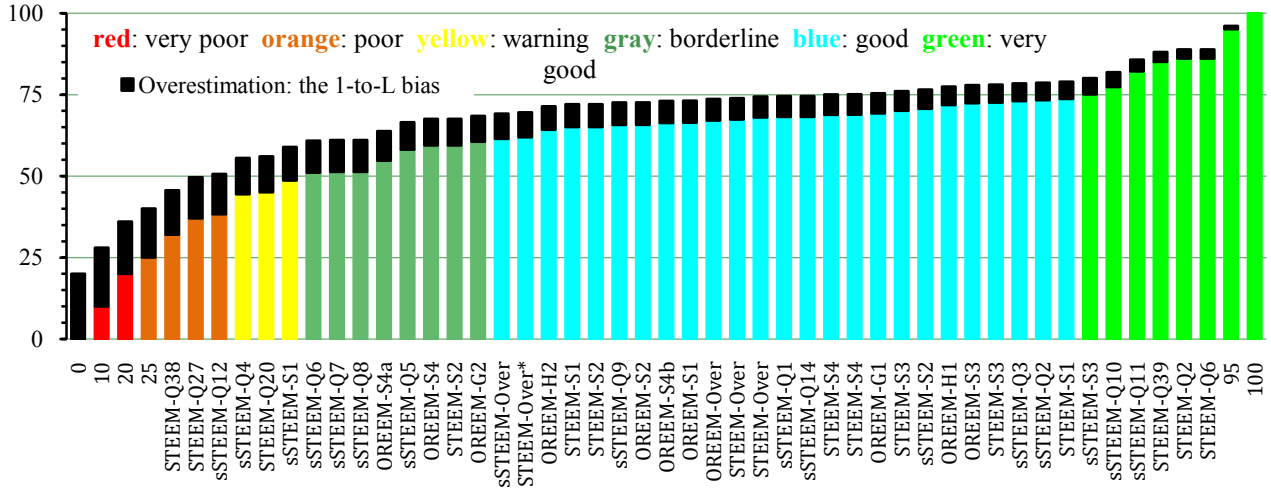
All standard mean scores were overestimated (Table 2, last column OE) by 2.8%–13.6% (mean 7.2%, median 6.7%). In addition, one in three of them (15/45 = 33%) had erroneously been sorted in an upper interpretation zone. Namely, almost one in four (11/45 = 24%) being in the 'good' zone had been interpreted as if they were in the 'very good' zone, and about one in ten (4/45 = 9%) being in the 'poor' zone had been interpreted as if they were in the 'good' one.

Table 3 presents the distribution in four interpretation zones (quarters) of all 45 reported non-0-based pseudostandard mean scores and the corrected 0-based standard ones. Almost three times more (17/6) pseudostandard than standard scores were found in the top interpretation zone (p=0.016). The high fraction of reported pseudostandard scores in the top quarter (38%, about thirteen times the DREEM equivalent from a recent review 3%; p=0.001) was eliminated after the appropriate correction (p=0.241).

Table 3. Distribution in four interpretation zones of the STEEM, OREEM and sSTEEM pseudostandard ($M_{n\%}$) and real standard ($M_{0\%}$) mean scores (n=45) and the DREEM standard overall mean scores (n=33) from a recent review [12]

Quarter; Interpretation Zone[1]	$M_{n\%}$	$M_{0\%}$	DR $M_{0\%}$
Worst (0–24.9); very poor	0 (0)	0 (0)	0 (0)
Second worst (25–49.9); poor	2 (4)	6 (13)	3 (9)
Second best (50–74.9); good	26 (58)	33 (73)	29 (88)
Best (75–100); very good	17 (38)	6 (13)	1 (3)
Total (0–100%)	45 (100)	45 (100)	33 (100)
Best versus the rest quarters (p value) [2]	---------- 0.016 ----------| ---------- 0.241 ----------|	
	---------- 0.001 ----------		

[1] In accordance with the DREEM interpretation guide [2].
[2] To meet chi-square conditions the three worst quarters were combined; p-values are based on chi-square test with Yates correction.



Separation of the 1-to-L bias (in black) from the real standard mean scores (their sum equals to the reported non-0-based pseudostandard mean scores). Disunion of the really good (blue: 60-74.9) from the borderline (gray: 50-59.9) scores, and the really poor (orange: 25-39.9) from the warning (yellow: 40-49.9) ones after Dimoliatis [12]. Data, in ascending order, from Table 2, plus six calculated scores using the formula in Figure 2 with x = 0, 10, 20, 25, 95, 100, in order to reveal what happens towards both edges.

Figure 3. Magnitude of the 1-to-L bias in comparison to the real standard mean score

4. Discussion

The STEEM, OREEM and student-STEEM non-0-based 1–5 question coding introduces an up to 20% overestimation of standard (percent) scores when assessing the quality of surgical educational environments that, to date, has escaped observation. The worse the quality of the environment, the greater the overestimation, beautifying things exactly where we need the warning bells to ring, i.e., in poor areas, especially in very poor ones. This reduces the usefulness of these otherwise very valuable instruments. Removing the 1-to-5 bias lead to this latent defect disappearing. This does not mean that other possibly coexisting biases in reporting [13] have also been eliminated. However, it does mean that there is no reason to believe that DREEM respondents have reporting biases different to those affecting STEEM, OREEM and sSTEEM responses. Surgical educational environment quality, as assessed by participants, appears to worsen after the 1-to-5 bias elimination; in reality, it had previously been erroneously overestimated.

A 0-based Likert scale should always be used when coding question response options, so that the most negative point would be coded as '0' [14], as originally recommended by Likert [15].

5. Conclusion

The non-0-based question coding in the STEEM, OREEM and student-STEEM questionnaires overestimates the quality of the educational environment due to the 1-to-5 bias or rather the 1-to-L bias, whereupon L indicates the number of points of the L-point Likert scale. Any non-0-based 'standard' score is a pseudostandard. The worse the educational environment the greater the overestimation is, beautifying things exactly when the alarm bell should be ringing. To raise the usefulness of these otherwise very good instruments, question coding should be always 0-based, i.e., the most negative point should be coded as '0', as originally recommended by Likert.

It is worth to note, that, generalizing, this should be applied to any Likert scale (L = 2, 3, 4 etc.) and any questionnaire of any field of study (education, quality of life, psychology, economics etc.), in order to avoid misleading statements or assumptions leading to inadequate political, economic, scientific or other related decisions.

REFERENCES

[1] S. Roff, S. McAleer, R.M. Harden, M. Al-Qahtani, A.A. Uddin, H. Deza, G. Groenen, P. Primparyon. Development and validation of the Dundee Ready Education Environment Measure (DREEM). Med Teach Vol. 19, No 4, 295–299, 1997.

[2] S. McAleer, S. Roff. A practical guide to using the Dundee Ready Education Environment Measure (DREEM). In J. M. Genn (ed), AMEE Medical Education Guide No 23 Curriculum, environment, climate, quality and change in medical education: a unifying perspective, Part 3. Association for Medical Education in Europe. Dundee UK, 2002

[3] S. Roff, S. McAleer, S. Skinner. Development and validation of an instrument to measure the postgraduate clinical learning and teaching educational environment for hospital-based junior doctors in the UK. Med Teach Vol. 27, 326–331, 2005.

[4] M.C. Holt, S. Roff. Development and validation of the Anaesthetic Theatre Educational Environment Measure (ATEEM). Med Teach Vol. 26, No. 6, 553–558, 2004.

[5] K. Cassar. Development of an instrument to measure the

surgical operating theatre learning environment as perceived by basic surgical trainees. Med Teach Vol. 26 No. 3, 260–264, (2004)

[6] J. Kanashiro, S. McAleer, S. Roff. Assessing the educational environment in the operating room – a measure of resident perception at one Canadian institution. Surgery Vol. 139 No. 2, 150–158, 2006.

[7] S. Nagraj, D. Wall E. Jones. The development and validation of the mini-surgical theatre educational environment measure. Med Teach Vol. 9, e192–e196, 2007.

[8] P. M. Fayers, D. Machin. Quality of life: assessment, analysis and interpretation. John Willey & Sons. West Sussex, England, ISBN 0-471-96861-7: pp 17, 141–142. 2000.

[9] RAND (2009). Scoring Instructions for MOS 36-Item Short Form Survey Instrument (SF-36). Available at: http://www.rand.org/health/surveys_tools/mos/mos_core_36 item.html (accessed 8 August 2013).

[10] A. Mahoney, P. J. Crowe, P. Harris. Exploring Australasian Surgical Trainees' Satisfaction with Operating Theatre Learning Using the 'Surgical Theatre Educational Environment Measure'. ANZ J Surg Vol 80, No 12, 884–889, 2010. doi: 10.1111/ j.1445-2197.2010.05430.x.

[11] M. R. Spiegel. Schaum's Outline of Theory and Problems of Probability and Statistics. McGraw–Hill, New York, Chapter 3: Theorems 3.1 to 3.7. 1995.

[12] I. D. K. Dimoliatis. The instrument Dundee Ready Education Environment Measure (DREEM) in Greek: how to use and preliminary results for the Greek medical educational environment. Archives of Hellenic Medicine, Vol. 27, No. 3, 509-521, 2010.

[13] D. L. Streiner, G. R. Norman. Health Measurement Scales - a practical guide to their development and use, 4th edition. Oxford University Press. Oxford, pp103–128, 143–151, 2008.

[14] R. A. Berk. Thirteen strategies to measure college teaching. Stylus publishing LLC, Sterling, Virginia, USA, p188, 2006.

[15] R. Likert. A technique for measurement of attitudes. Archives of Psychology Vol. 140, 44-53, 1932. In: R. A. Berk Thirteen strategies to measure college teaching. Stylus Publishing LLC, Sterling, Virginia, USA, p188, 2006.