Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 13, Number 8, September 2008 ISSN 1531-7714

Comparing Vertical and Horizontal Scoring of Open-Ended Questionnaires

Avi Allalouf, Galit Klapfer & Marina Fronton
National Institute for Testing and Evaluation (NITE)

Although horizontal scoring is generally considered more accurate than vertical scoring (due to the elimination of halo effects), no systematic comparison of the two methods had been carried out, prior to this study. Our extensive and structured study yields a comprehensive perspective on this issue. Our findings are that: (1) in general, there is not much difference between the methods; (2) horizontal scoring is somewhat better in terms of reliability and validity; (3) the raters' feedback pointed out the differences between the methods, with some in favor of one method, others in favor of the second method; and (4) the choice of scoring method makes a difference probably only with respect to a few specific questions.

Whereas the scoring of multiple-choice (MC) items is considered objective and highly reliable, the scoring of open-ended (OE) items (performance assessment, questionnaires…) has a subjective component; this kind of scoring is less reliable than MC, because OE involves human raters and is affected by their input. Nevertheless, a variety of means can be used in order to reduce the subjectivity inherent in the scoring of OE items and to improve its reliability. These means include: engaging professional raters, using comprehensive rating instructions, training the raters, monitoring the rating process, using retraining when drift is detected, having multiple raters, and engaging the services of an additional rater in cases of discrepancy between the raters. The last version of the "Standards for Educational and Psychological Testing" (AERA, APA & NCME, 1999) mentions briefly some of these topics.

Tests and questionnaires that consist of open-ended items can be scored in two ways: (1) vertically, where the rating unit is the whole test and the rater scores the entire test sequentially for each examinee in turn, and (2) horizontally, where the rating unit is a question and the rater scores the same question for a group of examinees before moving on to the next question.

Horizontal scoring is considered more accurate than vertical scoring because vertical scoring may suffer from a halo effect, i.e., the scoring of each particular question (excluding the first question) being dependent on the other (usually previous) questions to which the raters had already been exposed. (see, e.g., Rudner, 1992). Halo effects can occur even when the scoring instructions are very clear. The College Board AP Program (2006) justifies the use of horizontal scoring thus: "A reader could give an answer a higher or lower score than it deserves because the same student has performed well or poorly on other questions. To avoid this so-called 'halo effect,' in most cases each student's question is read by a different reader…."

However, vertical scoring is usually more practical and convenient. In horizontal scoring, the distribution and management of materials can be cumbersome, whereas in vertical scoring, methods are usually easier to manage.

Recently, Dore et al. (2006) conducted a study on the ABS (Autobiographical Submission) part of the admissions process for a medical school. One of the objectives of this study was to compare vertical and horizontal scoring. Their findings showed that (1) horizontal scoring has lower internal consistency (no halo effect, thus lowering internal consistency) and (2) horizontal scoring has higher inter-rater reliability (less non-relevant "noise" resulting from the halo effect). The conclusion, based also on correlations with another score used in the medical school admissions process, was

that horizontal scoring is preferable. However, although Dore et al. is a pioneering study, it has some major shortcomings: (1) it is based on too small a sampling to be reliable (30 examinees, eight questionnaire items, and two raters); (2) the Intraclass correlation was not used (it is considered the most suitable for estimating reliability); (3) there was only one validity indicator; (4) there was no item-level analysis (to examine whether some items are scored better horizontally and others vertically); and (5) there was no rater feedback. In our estimate, the question of which design – vertical or horizontal – is generally preferable has not been answered adequately. Dore et al. were aware of the limitations of their study and concluded: "… which method yields higher predictive validity … remains to be determined" (p. s72).

The objective of this study is to compare the psychometric characteristics of vertical scoring and horizontal scoring. The comparison was based on a standardized biographical questionnaire used in a medical school admissions process. The findings of this study are relevant for all manner of open-ended tests – such as performance assessment, for example.

METHOD

Instrument

A standardized Biographical Questionnaire (BQ) is one component of an assessment center that consists of a battery of tools for measuring non-cognitive attributes in candidates for medical schools (Gafni & Allalouf, 2005; Moshinsky & Rubin, 2005). The essential rationale for the BQ is that past behavior is a valid index for predicting future behavior. It contains open questions regarding (1) past experience and (2) emotional awareness; both types of questions are aimed at gauging a candidate's experience in coping with challenging emotional situations. Sample items of the BQ are presented in Appendix A.

The questionnaire contains 18 questions, each of which has a predefined objective. A detailed scoring rubric has been prepared. Most of the questions (11) have a score range of 0 – 5, two questions have a score range of 0 – 4, three questions have a range of 0 – 3, one question a range of 0 – 2, and one question a range of 0 -1. Candidates' replies are assessed by two different experts, and the final assessment is the average of these two evaluations. In cases of a substantial discrepancy between the two assessments, a third expert's assessment is included in the final score: the score is determined on the basis of the mean average of the third assessment and whichever of the original two is closest to it.

Examinees & Raters

180 medical school candidates (90 females, 90 males) were randomly sampled from among the candidates who took the BQ in 2005. Four experienced raters (two female, two male) evaluated the BQ in this study, after participating in an eight-hour preparatory workshop (as is the norm in operational scoring).

The operational scores of the 180 assessment center examinees (which were vertically scored) were gathered to serve as validity indicators.

STUDY DESIGN

The 180 questionnaires were randomly assigned to six groups, each of which consisted of 30 candidates (15 females, 15 males, in order to allow gender comparisons). Every questionnaire was rated by all four raters, twice vertically (as in operational rating) and twice horizontally. Every rater rated all 180 questionnaires, 90 vertically and 90 horizontally. The ratio of female and male raters was also balanced. Table 1 presents the study design.

Analysis

The two methods were compared on the test level and on the item level. Gender effects were analyzed as well. In addition, the four raters were asked to complete a feedback questionnaire regarding the two rating methods (see Appendix B).

RESULTS

Each of the 180 examinees received two final scores, one based on the vertical scoring, and one based on the horizontal scoring. Each score, at the test level and at the item level, is an average of the scores of the two raters. The score means were 39.9 (SD = 6.6) for the vertical scoring and 39.3 (SD = 6.4) for the horizontal scoring. The difference between scores was not statistically significant (paired samples t test, $p=0.15 > 0.05$), meaning that the expected halo effect was not observed here.

Table 1: Study Design

	Rater F = Female, M = Male	Group (Each group consists of 30 candidates, 15 females and 15 males)					
		A	B	C	D	E	F
Vertical Scoring	F1	●		●			●
	F2		●	●		●	
	M1	●			●	●	
	M2		●		●		●
Horizontal Scoring	F1		●		●	●	
	F2	●			●		●
	M1		●	●			●
	M2	●		●		●	

1. Test Level Analysis

Reliability

Two reliability measures were computed: 1) inter-rater reliability estimated by the Intraclass correlation coefficient, and 2) internal consistency reliability estimated by Cronbach Alpha. The results are presented in Table 2.

Intraclass correlation - The Intraclass correlation (Shrout and Fleiss, 1979) assesses reliability by comparing the variability of different ratings of the same subject to the total variation across all ratings and all subjects (See Appendix C).

Cronbach Alpha - estimates internal consistency. Internal consistency is expected to be greater in unidimensional tests and questionnaires. Since the BQ questionnaire is not perfectly unidimensional, very high estimates are not expected (but medium estimates definitely are).

Although the results indicate a slight advantage for the horizontal method, the difference is not statistically significant. The median correlation between the vertical scores and the horizontal scores was very high, 0.90, indicating that there is not much difference between the two scoring methods.

Table 2: Inter-Rater Reliability[1] and Internal Consistency[2], by Group

Reliability	Scoring	Group								
		A	B	C	D	E	F	**Mean**	**Median**	**Reliability**[3]
Inter-Rater Correlation	Vertical	0.76	0.54	0.69	0.78	0.92	0.66	**0.73**	**0.73**	**0.84**
	Horizontal	0.82	0.76	0.71	0.88	0.68	0.75	**0.77**	**0.76**	**0.86**
Internal consistency	Vertical	0.64	0.57	0.66	0.68	0.69	0.68	**0.65**	**0.67**	-
	Horizontal	0.70	0.71	0.70	0.70	0.66	0.72	**0.69**	**0.70**	-

1 Based on Intraclass correlation
2 Cronbach Alpha
3 After applying a Spearman- Brown correction to estimate reliability for two raters based on the reliability of one rater.

Table 3: Correlations between Vertical and Horizontal Scores and Operational Scores from the Assessment Center (N=180)

	BQ[1]	JDQ[2]	SS[3]	ZR[4]
Vertical Scoring	0.80	0.16	0.27	0.47
Horizontal Scoring	0.82	0.28[5]	0.28	0.51

1 Biographical Questionnaire (operational)
2 Judgment and Decision-Making Questionnaire – examination of the candidates' ability to contend with complex situations and moral dilemmas
3 Simulation Stations – observation of candidates' behavior in simulation and debriefing stations
4 Final score, based on the following weights BQ - 1, JDQ - 1, SS - 3
5 The difference between 0.28 and 0.16 is statistically significant

Validity Indicators

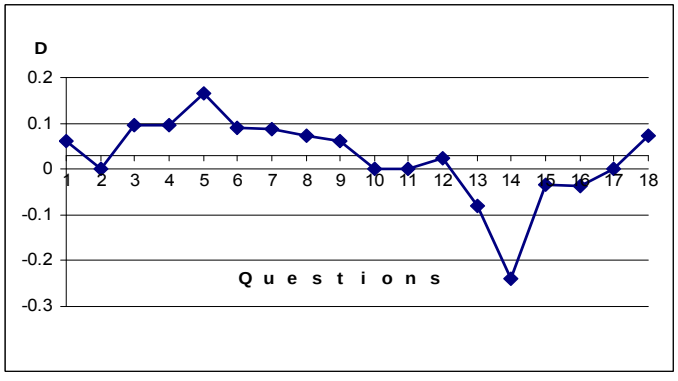
Correlations with the operational BQ score (vertically scored) and with other components of the assessment center served as validity indicators. Table 3 presents these correlations. In all comparisons, the horizontal scoring has somewhat higher correlations, but only in the JDQ score is the difference statistically significant.

2. Item-Level Analysis

Score Differences

The *standard mean difference* D ($= \frac{\bar{x} - \bar{y}}{\sqrt{(s_x^2 + s_y^2)/2}}$) between groups was calculated for each question (x – vertical scoring, y – horizontal scoring). Figure 1 presents the Ds for the 18 questions.

Figure 1: Standard Mean Differences (D) Between Vertical & Horizontal Scoring, by Question
($\overline{D}$ = 0.026, appears in the figure)



The average D is very small, 0.026, meaning that, on average, no difference was found between the two methods. In two items (5 & 14) the difference, though small, is statistically significant. Looking at the content of the items for which there was a relatively higher difference between the two scoring methods (i.e., item 14 according to Table 4) did not contribute to our understanding of the causes for these differences. However, it should be noted that statistically, one or two divergent items are to be expected from among the eighteen items even if all eighteen are suitable for both rating methods.

Agreement Indices

Table 4 presents the agreement and adjacent agreement by item and rating design. It demonstrates that the agreement is an attribute of the question rather than of the scoring method. The agreement and adjacent agreement means of the two rating methods are very similar. It is also evident from the table that the scoring method usually makes no difference. The correlation between the agreement levels is high, 0.96 for the agreement, 0.78 for the adjacent agreement (the last is lower, due to smaller variance of the variables)

Kappa Correlations

Kappa (Cohen, 1960) quantifies the *level of agreement* using the proportion of chance (or expected) agreement. Kappa compares the actual agreement to the proportion of times raters would agree by chance alone (See Kappa and weighted Kappa in Appendix C). In our data, we

Table 4: Agreement and Adjacent Agreement by Item and Rating Design

Item	Range	Agreement		Adjacent Agreement	
		Vertical	Horizontal	Vertical	Horizontal
1	0-3	88%	90%	100%	99%
2	0-2	97%	97%	100%	100%
3	0-5	64%	71%	91%	89%
4	0-5	53%	58%	86%	89%
5	0-5	46%	49%	81%	82%
6	0-5	32%	41%	82%	86%
7	0-5	45%	52%	81%	92%
8	0-5	46%	44%	80%	83%
9	0-5	42%	39%	83%	83%
10	0-1	98%	100%		
11	0-5	51%	47%	91%	91%
12	0-5	54%	53%	86%	82%
13	0-5	34%	36%	84%	75%
14	0-4	66%	53%	93%	82%
15	0-4	62%	58%	93%	88%
16	0-3	62%	69%	95%	94%
17	0-3	52%	51%	97%	96%
18	0-5	37%	47%	82%	76%
Mean		**57%**	**58%**	**89%**	**87%**
SD		20%	19%	7%	7%

Table 5: Rater Reactions[1]

Variable	Scoring Method	
	Vertical	Horizontal
Halo effect	There is halo effect, but in some instances, the answers to previous questions help	No halo effect (+)
Speed	Vertical scoring is slower	Faster (+)
Level of fatigue	Vertical scoring is less tiring (+)	More tiring
Enable rater feedback to scoring instructions	In vertical scoring, the raters' feedback on the scoring instructions for a specific question is less immediate	Here, the raters can provide immediate and efficient feedback (+)
Provide an overall impression of the candidate	It is possible to provide an overall impression of the candidate (+)	Not possible
Getting used to handwriting & style	The rater gets familiar with the handwriting & style of the examinee (+)	Not possible

1. + Indicates an advantage of the design

computed the weighted Kappa for all the items. The mean weighted Kappa for the vertical scoring and the horizontal scoring were found to be very close (0.54 and 0.55, respectively).

3. Gender Effects

Gender effects were analyzed; each of the examinees received two final scores, one based on the vertical scoring, and one based on the horizontal scoring. For the vertical scoring the mean scores were 39.4 for the males, and 40.3 for the females; for the horizontal scoring the mean scores were 38.8 for the males and 39.9 for the females. The differences in scores according to gender were very similar for each rating, meaning that no gender effect was identified.

4. Rater Reactions

The four raters were given a questionnaire to complete regarding the two rating designs. Their reactions are summarized in Table 5. These reactions are very important as they "put on the table" a few important variables in addition to the halo effect.

Horizontal scoring is faster, enables the raters to provide immediate and efficient feedback to the professionals who are responsible for the rating process, and, of course, has no halo effect. On the other hand, vertical scoring is less tiring, provides an overall impression of the examinee and allows one to get used to his or her handwriting and style. Regarding the halo effect, according to the raters, in some cases, the examinee answers to previous questions help them in rating a specific question. Overall, taking all the variables into account, neither of the two designs appeared to be better.

DISCUSSION

Although horizontal scoring is generally considered more accurate than vertical scoring (due to the elimination of halo effects), no systematic comparison of the two methods had been carried out, prior to this study. A recent study devoted to the subject (Dore et al., 2006) did not provide an adequate treatment of it. Our extensive and structured study yields a comprehensive perspective on this issue. Our findings are that:

1. in general, there is not much difference between the methods;
2. horizontal scoring is somewhat better in terms of reliability and validity;
3. the raters' feedback pointed out the differences between the methods, with some in favor of one method, others in favor of the second method;
4. the choice of scoring method makes a difference probably only with respect to a few specific questions. This is perhaps because of the clear and precise scoring instructions adhered to by the raters, which minimize the halo effect in vertical scoring.

Application of horizontal scoring requires that the candidates be informed of the rating method used, so that they answer each question independently, i.e., without reference to previous questions. Moreover, it is essential to specify the number of questions a rater should rate consecutively, since horizontal scoring is a tiring process.

Our findings, taken in conjunction with a number of logistical considerations, do not necessarily support use of the horizontal scoring method. As long as the vertical scoring method is easier to employ, there is no need to replace it with horizontal scoring, which is less practicable.

The issue of vertical vs. horizontal scoring can be applied to an open test that contains some open items and to a personal questionnaire which serves as a standardized written version of a structured interview. A halo effect that occurs on an open test can cause the rater to score the examinee higher or lower in one item, based on the quality of his/her answer to previous items; a halo effect that occurs on a biographical questionnaire is somewhat different. Previous answers sometimes indicate relevant information regarding the examinee, which tends to result in a higher examinee score.

In this kind of questionnaire, and probably in open-ended tests in general, it is sometimes necessary to read a sequence of several questions together in order to fully understand and rate the answers. Therefore, one might also consider applying vertical rating to groups of questions. The scoring method should be adjusted to suit the specific questionnaire – basically, the horizontal method is preferable, yet, some questions could be scored as a "vertical group." This idea is similar in some aspects to the item bundle model. (see Rosenbaum, 1988).

Further research on this topic is needed. We recommend studying the following: a) applying factor

analysis in each scoring method and comparing the factor structures, b) expanding the gender effect analysis, including interaction between rater & examinee, c) repeating the study with naïve, untrained raters, and d) repeating the study with other kinds of tests, such as achievement assessments.

REFERENCES

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, D.C.: AERA.

Cicchetti, D.V. & Allison, T. (1971), A new procedure for assessing reliability of scoring EEG sleep recordings, *American Journal of EEG Technology* 11,101-109.

Cohen, J. A. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20, 37-46

The College Board AP Program (2006) http://apcentral.collegeboard.com/ Home > The Exam > All About the Exams > Exam Scoring

Dore, K. L., Hanson, M., Reiter, H. I., Blanchard, M., Deeth, K. & Eva, K. W. (2006)

Medical School Admissions: Enhancing the Reliability and Validity of an Autobiographical Screening Tool. *Acad. Med.* 81 70-73.

Gafni, N. & Allalouf, A. (2005, April) *Score Analysis of the Non-Cognitive Measures Used in Student Selection for Medical Schools.* Paper presented at the annual meeting of the American Educational Research Association

Maclure, M. & Willett, W.C. (1987). Misinterpretation and misuse of the kappa statistic. *American Journal of Epidemiology* 126, 161-169.

Moshinsky, A. & Rubin, O (2005, April). *The Development and Structure of an Assessment Center for the Selection of Students for Medical School.* Paper presented at the annual meeting of the American Educational Research Association

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika* 53, 349-539.

Rudner, Lawrence M. (1992). Reducing errors due to the use of judges. *Practical Assessment, Research & Evaluation*, 3 (3).

Shrout, P.E. & Fleiss, J.L. (1979) Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin* 2, 420-428.

APPENDIX A.
BQ - Sample questions

1. Do you engage in regular leisure activities (hobbies)?

_____

_____

2. Do you intend to continue these activities during your studies?

_____

_____

3. Do you think you will manage to combine studying with these leisure activities?

_____

_____

4. In your opinion, what does the statement "the patient has the final say" mean? How should you act when this principle is not in keeping with the patient's best interests?

_____

_____

APPENDIX B
Raters Questionnaire

Name of Rater: _____ Date: _____

-- Rater Questionnaire --

There are two rating methods:
1. Vertical rating – where the rating unit is the entire questionnaire. Each rater receives a questionnaire and evaluates it in its entirety before moving on to the next questionnaire.
2. Horizontal rating – where the rating unit is the individual question. Each rater receives a single question to evaluate and only moves on to the next question when the initial one has been evaluated by all raters.

A. With regard to the Biographical Questionnaire

Vertical rating

Pros: _____

Cons: _____

Horizontal rating

Pros: _____

Cons: _____

B. With regard to the Dilemmas

Horizontal rating (the current practice)

Pros: _____

Cons: _____

Vertical rating (evaluating all three dilemmas for each examinee consecutively)

Pros: _____

Cons: _____

C. Which method is preferable for each type of material, in your opinion?

Biographical Questionnaire _____ Dilemmas _____

d. Do you have additional comments you wish to make in this regard?

_____

Thank you for your cooperation.

APPENDIX C
Intraclass and Kappa Correlations

The Intraclass Correlation (ICC) assesses ratings reliability by comparing the variability of different ratings of the same subject to the total variation across all ratings and all subjects. Intraclass is used to measure inter-rater reliability for two or more raters. Shrout and Fleiss (1979) describe three classes of ICC reliability; each relates to a different rater agreement study design. These three cases correspond to the standard ANOVA models. In our study we use *Case 2*, which corresponds to the two-way ANOVA random-effects model. It relates to a random sample of k raters selected from a larger population; in this case, each of the examinees is rated by each rater.
Intraclass correlation takes into account the variance between raters and the variance between examinees.
The formula for *Case 2*

$$r_2^2 = \frac{MS_{ex} - MS_{res}}{MS_{ex} + (n_r - 1) * MS_{res} + (\frac{n_r * (MS_r - MS_{res})}{n_{ex}})}$$

- mean square effect for examinees MS_{ex}
- mean square residual effect MS_{res}
- mean square raters effect MS_r

The Kappa Correlation (Cohen, 1960). One of the possible uses of Kappa is as a way of quantifying the *level of agreement* (i.e., as an effect-size measure). Kappa's calculation is based on the proportion of chance (or expected) agreement. This is interpreted as the proportion of times raters would agree by chance alone compared to the actual agreement. The term is relevant only under conditions of a statistical independence of raters. With ordered category data, one must select weights arbitrarily to calculate weighted kappa (Maclure & Willet, 1987).

The weighted kappa coefficient is a generalization of the simple kappa coefficient, using weights to quantify the relative difference between categories.

The weights w_{ij} are constructed so that $0 \le w_{ij} < 1$ for all $i \neq j$, $w_{ii} = 1$ for all i, and $w_{ij} = w_{ji}$. The weighted kappa coefficient is defined as

$$\hat{k}_w = \frac{P_{O(w)} - P_{E(w)}}{1 - P_{E(w)}} \quad \text{while} \quad P_{O(w)} = \sum_i \sum_j w_{ij} p_{ij} \;, \quad P_{E(w)} = \sum_i \sum_j w_{ij} p_{i\bullet} p_{\bullet j}$$

The SAS statistical program computes kappa coefficient weights from the Cicchetti-Allison (1971) formula:

$w_{ij} = 1 - \frac{|C_i - C_j|}{C_c - C_1}$ where C_i is the score for column i, and C is the number of categories or columns.

Citation

Allalouf, A., Klapfer, G. & Fronton, M. (2008). Comparing Vertical and Horizontal Scoring of Open-Ended Questionnaires. *Practical Assessment Research & Evaluation*, 13(8). Available online: http://pareonline.net/getvn.asp?v=13&n=8

Authors

Avi Allalouf, PhD
Director of Scoring & Equating
National Institute for Testing and Evaluation (NITE)
PO Box 26015
Jerusalem 91260 Israel

Email: avi [at] nite.org.il

Galit Klapfer
National Institute for Testing and Evaluation (NITE)
PO Box 26015
Jerusalem 91260 Israel

Marina Fronton
National Institute for Testing and Evaluation (NITE)
PO Box 26015
Jerusalem 91260 Israel