Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 16, Number 9, June 2011 ISSN 1531-7714

Too Reliable To Be True? Response Bias as a Potential Source of Inflation in Paper-And-Pencil Questionnaire Reliability

Eyal Peer, *Hebrew University of Jerusalem* and
Eyal Gamliel, *Ruppin Academic Center, Israel*

When respondents answer paper-and-pencil (PP) questionnaires, they sometimes modify their responses to correspond to previously answered items. As a result, this response bias might artificially inflate the reliability of PP questionnaires. We compared the internal consistency of PP questionnaires to computerized questionnaires that presented a different number of items on a computer screen simultaneously. Study 1 showed that a PP questionnaire's internal consistency was higher than that of the same questionnaire presented on a computer screen with one, two or four questions per screen. Study 2 replicated these findings to show that internal consistency was also relatively high when all questions were shown on one screen. This suggests that the differences found in Study 1 were not due to the difference in presentation medium. Thus, this paper suggests that reliability measures of PP questionnaires might be inflated because of a response bias resulting from participants cross-checking their answers against ones given to previous questions.

Self-reporting questionnaires are frequently used to measure educational and psychological variables. However, such questionnaires raise concerns about the presence of a response bias. This bias is defined as "a systematic tendency to respond to a range of questionnaire items on some basis other than the specific item content" (Paulhus, 1991).

Researchers have split response bias into two broad categories: response style and response set (Paulhus, 1991). *Response style* is the tendency to distort responses in a particular direction, more or less regardless of the content of the stimulus. *Response set* is the conscious or unconscious desire on the part of the respondent to answer in such a way as to produce a certain picture of oneself. Researchers have suggested that responding in a desirable way is a response set, which is a situational and temporary response pattern. In contrast, response style is a more long-term trait-like quality that is assumed to remain similar across different questionnaires (see Paulhus, 1991, for a detailed review).

The literature details several examples of response sets: *Random responding* is a response set where participants answer questions with little pattern or thought (Cronbach, 1950; Osborne & Blanchard, 2011); *Malingering* refers to participants falsifying their answers in order to present themselves in more negative light (Osborne & Blanchard, 2011); *Dissimulation* refers to participants altering their answers in order to achieve certain goals, for example social desirability – conforming to social norms in order to "look good" (e.g., Bardwell, Ancoli, & Dimsdale, 2001; McKelvie, 2004; Sullivan & Scandell, 2003).

Two of the proposed methods to reduce the effect of response bias, whether response set or response style, include scrambling the questions' order (e.g., Ruble & Stout, 1990, 1991), and reversing the scale of some questions such that high-scale values reflect a low value in the measured attribute (e.g., Tibbles, Waalen, & Hains, 1998). However, these attempts obviously do not

eliminate the effect of response set or response style on participants' answers.

Response bias may also emerge when individuals' responses to items are affected by their responses to preceding items. Such response bias could serve several purposes: participants might keep a positive image of themselves as consistent (and possibly rational) or it might help them to be more quick and efficient in completing the questionnaire by "copying" their previous answers [1]. Whatever the causes of such response bias are, it leads to artificially consistent responses and inflated internal consistency. Such an artificial increase of internal consistency might also be caused by response sets as malingering and dissimulation. In contrast, artificial decrease of internal consistency would follow response set of random responding (Osborne & Blanchard, 2011).

The growing use of computerized and Internet-based questionnaires for measuring educational and psychological variables opens new avenues for examining response bias. Computerized questionnaires offer various advantages over paper-and-pencil (PP) questionnaires (e.g., Buchanan, 2002; Gosling, Vazire, Srivastava, & Oliver, 2004). Several studies examined the psychometric qualities of computerized questionnaires, either independently (e.g., Fanciullo, Jamison, Chawarski, & Baird, 2003; Kleiman & Gati, 2004; McCue, Martin, Buchanan, Rodgers, & Scholey, 2003) or by comparing them to traditional PP questionnaires (e.g., Mertler & Earley, 2002, 2003; Miller et al., 2002; Riva, Teruzzi & Anolli, 2003; Whittier, Seeley, & St. Lawrence, 2004). These studies typically concluded that the mode used is immaterial in terms of the psychometric properties of the questionnaires.

One central psychometric property habitually examined in this context is internal consistency. Internal consistency is probably the most frequently used reliability measure in psychological and educational research, and the most popular index of internal consistency is Cronbach's (1951) coefficient alpha (Schmidt, Le, & Ilies, 2003). Cronbach (1951) presented coefficient alpha in two manners – a conceptualized manner and a computational manner. Conceptually, alpha is "the mean of all split-half coefficients resulting from different splitting of a test. […] alpha is therefore an estimate of the correlation between two random samples of items from a universe of items like those in the test" (p. 297). Cronbach presented a formula to calculate alpha as:

$$\alpha = \frac{n}{n-1}(1 - \sum_i V_i / V_t) \ ,$$

where n is the number of items, V_i is the variances of items i=1 to I, and V_t is the variance of the total score (Cronbach, 1951, p. 299). Thus, whenever there is no internal consistency between the items, that is, the correlation between the items is zero, the covariance between the items is zero, and the sum of item variances equals the variance of the total score; in such case, the formula would yield a zero result suggesting no internal consistency between the items. As the internal consistency (i.e., the correlation and the covariance coefficients) increase, the alpha would increase, until the extreme case of full consistency: perfect correlations between items, yielding a sum of item variances that exceeds substantially the total variance, yielding a close to zero ratio between V_i and V_t; in such a case, the coefficient would yield a value of 1.

It should be noted that Cronbach's alpha is a measure of internal consistency of items comprising a test or a self-reported questionnaire. As such, alpha is affected by measurement error causing inconsistency between items, but it is not affected by other sources of measurement errors, such as the participants' physiological and psychological state, the situation and context of administering the test/questionnaire, and the examinee/rater. Indeed, generalizability theory enables to differentiate between several errors of measurement that correspond to different true scores. Of the three sources of error measurement dealt with by classical test theory – trait stability over time, domain or content sampling, and item variability, internal consistency (e.g., Cronbach's alpha) is affected only by the last one (Rodriguez & Maeda, 2006).

Several studies compared the internal consistency of computerized (or online) questionnaires to their respective PP questionnaires, as measured by Cronbach's alpha coefficient (e.g., Potosky & Bobko, 1997; Mertler, 2003). With respect to self-reported educational and psychological variables, experimental studies reported similar alpha coefficients for web-based questionnaires compared to printed copies of the same questionnaires (respective alphas .88-.91 vs., .88-.89; Mertler & Earley, 2002, 2003). Other experimental

[1] We thank an anonymous reviewer for this suggestion.

studies found somewhat higher values for PP questionnaires (.83 and .84) relative to their online versions (.75 and .74, respectively; Riva et al., 2003).

Comparing PP questionnaire internal consistency to computerized versions of the same questionnaires should help in examining response bias. One study tried the following approach: using a computerized questionnaire, once participants responded to an item, the response scale window of the item was minimized. This prevented participants from seeing their answers to already completed questions, as they proceeded through the questionnaire. This resulted in somewhat lower reliability measures compared to PP versions of the same questionnaire (Gamliel & Davidovitz, 2005). This finding suggests that computerized versions of questionnaires can help in revealing such instances of response bias.

Other ways of circumventing or revealing response bias could be, for example, presenting items on separate screens when using computerized versions, instead of presenting all items simultaneously on one page. The same can be done with traditional PP questionnaires, though the procedure is much more cumbersome. Moreover, the use of a computerized questionnaire enables simple manipulation of the visual presentation of both items and scales. For instance, the computerized questionnaire can present a small number of items simultaneously. Using such techniques might reduce response bias by hindering participants' attempts to rely on answers to previous items, or on the visual pattern of their answers that is visible when using PP questionnaires. This predicted reduction in response bias is expected to result in lower measures of internal consistency for the computerized versions of questionnaires.

In the current study, we focused on manipulating the visual presentation of the number of questions presented per screen in a computerized questionnaire. We then compared the resulting internal consistency measures to the one obtained in an equivalent PP questionnaire. We theorized that the mode of presentation would cause artificially homogeneous responses in the PP version, possibly due to a response bias resulting from participants cross-checking their answers against earlier responses. This, of course, would reduce the variability of the responses. We hypothesized that the internal consistency of the traditional PP questionnaire—when all items and their corresponding scales appear together on one page—would be higher than that achieved in computerized versions presenting only some items per screen.

Study 1

Method

Participants. The participants were 287 undergraduate students, 23 males and 259 females (5 participants did not state their gender). Participants' mean age was 25.0, with a standard deviation of 4.0.

Materials. The participants filled out the short version of the Need for Cognition (NC) questionnaire, consisting of 18 items that measures the individual tendency to engage and to enjoy in effortful cognitive endeavors (Cacioppo & Petty, 1982). This scale was chosen because it is relatively short and has sound psychometric properties including high reliability and validity (Cacioppo, Petty, & Kao, 1984). Participants stated the degree of their agreement with each item on a 5-point scale ranging from minimal agreement (1) to maximal (5). Items number 3, 4, 5, 7, 8, 9, 12, 16 and 17 were reverse coded in order to maintain consistent scale direction for all items.

Design and Procedure. All participants filled out the questionnaire in a computer laboratory, in which one or two participants were seated at a time, carefully monitored by an experimenter. Participants were randomly assigned to one of four conditions: Condition 1 – Participants filled out the PP version of the questionnaire, in which all 18 items were presented on a single page; Conditions 2-4 – Participants filled out the questionnaire on one of the lab computers. Conditions 2-4 varied in the number of items that were simultaneously presented on screen: one, two or four items simultaneously shown on each screen. The order of the items in all conditions followed the PP version's order. Once a participant responded to the question(s) on a screen, she clicked the "continue" button and the next question(s) were presented (participants could not go back to previous screens). Other than this, all other features of the computerized questionnaire mirrored the PP version including the wording of the items and the scale labels.

The research was presented to all participants as a study of people's personal thinking style. The students were told that there were no correct or incorrect answers since different people describe themselves differently. All questionnaires were anonymous and all participants

were told that the information obtained would be used for research purposes only.

Results and discussion

We hypothesized that internal consistency would be affected by the number of items shown to participants and would be highest in the PP questionnaire (when all questions are visible) and lowest when only one item is presented on each screen.

The following results were computed only for participants who filled out all 18 items in the questionnaire. The total NC index was calculated as the average of responses to all 18 items. The averages of the NC index for the four conditions were very similar – ranging between 3.59 and 3.74. The different conditions were related to less than 1% of the total variance of the NC index and the between-groups' differences were not statistically significant, F (3,283) = 1.05; p = .37.

Table 1 presents descriptive statistics and Cronbach's (1951) alpha measure for internal consistency. The differences in Cronbach's alpha coefficients for the four conditions clearly show that the internal consistency was higher in the PP questionnaires (about .90) relative to the same questionnaire presented on a computer screen with one, two or four questions per screen (alpha measures of .83, .68 and .79, respectively). In order to examine whether these differences were statistically significant, we used Hakstian & Whalen's (1976) test for significance of the differences between independent Cronbach's alpha coefficients. The test produced a statistically significant result (Chi square (df=3) = 27.31, p < .001) that supported the above conclusion[2].

Thus, the hypothesis that the PP questionnaire would have a higher internal consistency than computerized versions of the same questionnaire was supported. This may be regarded as evidence of a response bias in participants' responses to the PP questionnaire: participants may have inadvertently modified their responses to certain items so that they would correspond to previous items. Since in the computer versions participants only saw their responses to some items per screen, and not all of their responses to all the items thus far answered, the response bias in these conditions was reduced, as was the internal consistency of the questionnaire in these conditions. The Cronbach's alpha coefficient of the one question per screen condition deviates from the expected reverse monotone relation between the number of items per screen and the internal consistency measure. We will address this result and try to offer explanation for it in the General discussion section.

Table 1: Descriptive statistics and internal consistency measures for the experimental conditions in Study 1 and Study 2

Condition	N	Mean	sd	alpha
Study 1				
Paper-and-pencil	90	3.63	0.56	.91
One question per screen	86	3.71	0.49	.83
Two questions per screen	56	3.74	0.33	.68
Four questions per screen	55	3.61	0.42	.77
Study 2				
Paper-and-pencil	43	3.58	0.52	.90
One question per screen	51	3.67	0.39	.74
All questions on one screen	25	3.70	0.47	.88

One may, however, argue that it was not the number of questions presented on each screen that caused the differences in the internal consistency between the versions. Rather, the differences may be attributed to the difference in the medium of presentation: PP vs. computerized. Perhaps the computerized versions had a lower reliability than the PP version and the differences observed were not due to the number of questions presented on each screen and so do not denote the occurrence of a response bias in the PP version.

There are, at least, two ways to empirically test this argument. The first is to manipulate the number of questions presented without changing the medium of presentation. Namely, to use a PP version and to manipulate the number of questions presented on each page so that in some conditions only some of the questions are shown on each page. The other method is to create a duplicate of the PP version on a computer screen by presenting all questions on a single screen, the same way as they are presented on a single page. Because the application of the second method is less cumbersome and offers more control on what participants do, we used this method in devising Study 2.

[2] Because the scale had 18 items, the last screen of the condition presenting four items per screen presented only two items (no. 17 and 18). Re-analysis of the data obtained for the first 16 items of the scale found similar results for the four experimental conditions.

Study 2

Method

Participants. The participants were 119 undergraduate students, 11 males and 102 females (6 participants did not state their gender). Participants' mean age was 24.5, with a standard deviation of 3.7.

Materials. As in Study 1, participants filled out the short version of the Need for Cognition (NC) questionnaire, consisting of 18 items (Cacioppo et al., 1984). Participants stated the degree of their agreement with each item on a 5-point scale ranging from minimal agreement (1) to maximal (5).

Design and Procedure. All participants filled out the questionnaire in a computer laboratory. Participants were randomly assigned to one of three conditions. Conditions 1 and 2 followed Conditions 1 and 2 of Study 1 (PP and one question per screen, respectively). In condition 3 participants were presented with a full screen of the questionnaire, which was an exact duplicate of the PP version. Participants were able to see all the questions on one screen and for each question they were instructed to tick the box underneath the number that best describes their level of agreement with the statement. All other instructions and setting followed Study 1.

Results and discussion

No statistically significant differences were found for the NC score in the three conditions, F (2,116) = 0.66; p = .52, η^2=.01. In contrast, as can be seen in Table 1, the reliability of the PP version was again higher than that of the computer version with one question per screen (Condition 2) (.9 vs. .74, respectively), but was not much different from the reliability of the full screen version (.88). As in Study 1, Hakstian & Whalen's (1976) test produced a statistically significant result (Chi square [df=2] = 10.07, p < .01) that showed that the differences between the internal consistency of the different conditions were statistically significant. This suggests that while showing only one question per screen reduced the questionnaire reliability, the medium itself was not the cause. This finding supports the conclusion that the differences found in Study 1 were not due to the differences in the medium of presentation but were the result of the differences in the number of questions presented in the various conditions.

General discussion

This study showed the effect of response bias on the internal consistency of a questionnaire by manipulating the number of items presented simultaneously. When all items were presented simultaneously using the traditional PP method, the internal consistencies tended to be higher than the ones obtained by presenting fewer items (one, two or four) separately on a computer screen. These findings replicate previous results of studies that compared PP vs. computerized versions of questionnaires (Mertler & Earley, 2002, 2003). Moreover, the similar consistency coefficients found in Study 2 between the PP questionnaire and when all questions were shown on a single computer screen, suggests that the mode (paper vs. computer-based) did not intrinsically matter. Rather, the differences found in Study 1 between the reliability of PP questionnaire and computerized versions with some items per screen were probably the result of a response bias: Participants were more consistent in their responses when all items were visible.

Nevertheless, we did not find a linear relation between the number of items presented and the internal consistency of the questionnaire. This may be due to the fact that some items were reversely coded. It is possible that these items also affected the internal consistency of the questionnaires in the various versions, and that this effect confounded the effect of the number of items presented on each screen. Indeed, previous research suggested that reversely coded items might produce artificial factor, in addition to the one that was originally measured by the other items (Spector., Van Katwyk, Brannic, & Chen, 1997). Future research should manipulate the number of items reversely coded, or use other measures to control for the possible effect of this factor.

This study suggests that internal consistency measures typically reported for traditional PP questionnaires measuring educational and psychological variables may have been artificially inflated. Additional research is needed in order to confirm this suggestion, using various manipulations to prove (or disprove) that it is indeed the number of visible items that causes the differences in reliability. Although two studies presented in this paper yielded statistically significant results confirming the hypothesis, there is a need for future replication. One reason is that the alpha measure is highly dependent on the particular sample used. For

example, the experimental condition "one question per screen" yielded coefficients of .74 and .83 in the two studies. In addition, the null hypothesis statistical significance technique for statistical inference has many limitations (e.g., Cohen, 1994). Thus, as it was previously suggested, "given the problems of statistical induction, we must finally rely, as have the older sciences, on replication" (Cohen, 1994, p. 1002).

Future research could follow several directions: First, it is important to explore more variations in the number of items appearing together. Second, this phenomenon should be replicated using different questionnaires with either more or less questions with some reverse coded and others not. Third, experimental manipulations can be made to examine how and to what extent people base their responses to answers on answers they have already given to previous items. For example, a computerized experiment can emphasize previous responses by displaying them on part of the screen or by displaying a mean score of previous responses. This kind of manipulation is hypothesized to increase response bias and internal consistency. On the other hand, one can try to reduce response bias by obscuring previous responses. One method – dividing the questionnaire into single items on different screens – was employed in the current study. Another method could be to show all items but remove an item once it has been marked. Finally, putting participants under a cognitive load (by asking them to perform another task simultaneously) might also hinder attempts to answer items so that they correspond with previous items and thus reduce internal consistency of the questionnaires.

If future research does indeed confirm the suggested artificial inflation of PP questionnaire internal consistency, several implications are worth mentioning. First, the validity coefficients of variables measured using traditional PP questionnaires, although they are probably measured with higher internal consistency, are not expected to be higher compared to the validity coefficients of variables measured using computers. Although validity is a function of reliability (e.g., Crocker & Algina, 1986), the higher values of internal consistency of the former measures are artificially inflated, and are not expected to contribute to the validity of these measures.

Secondly, computerized and online questionnaires would gain a paradoxical advantage – their relatively low internal consistency measures. This apparent disadvantage of computerized and online questionnaires is actually an advantage, because the higher internal consistency coefficients of PP questionnaires are artificially inflated.

In addition, researchers measuring educational and psychological variables would be faced with a dilemma: if they use typical PP questionnaires, presenting all items and their response scales simultaneously, the internal consistency would be higher than if fewer items were presented simultaneously and/or the response scales were hidden. On the one hand, external, non-scientific factors might convince the researcher to use the former version of presentation, and so the likelihood of obtaining high internal reliabilities, although probably not high validity coefficients. On the other hand, scientific factors might convince the researcher not to take advantage of the artificially inflated high internal consistency and to present fewer items simultaneously and/or hide the answers to previous questions when using computerized questionnaires. Lastly, when computerized applications of the questionnaire are not available, it is important to devise a method to estimate the effect of response bias on the reliability of PP questionnaires and perhaps help correct the ratings obtained in this method.

References

Bardwell, W. A., Ancoli, I. S., & Dimsdale, J. E. (2001). Response bias influences mental health symptom reporting in patients with obstructive sleep apnea. *Annals of Behavioral Medicine, 23*(4), 313-317.

Buchanan, T. (2002). Online assessment: Desirable or dangerous? *Professional Psychology: Research and Practice, 33*(2), 148-154.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*, 116-131.

Cacioppo, J. T., Petty, R. E., & Kao, C. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48*, 306-307.

Cohen, J. (1994). The earth is round (p<. 05). *American Psychologist, 49*(12), 997-1003.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement, 10*, 3-31.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.

Fanciullo, G. J., Jamison, R. N., Chawarski, M. C., & Baird, J. C. (2003). Reliability and validity of an interactive computer method for rating quality of life. *Pain Medicine, 4*(3), 257-268.

Gamliel, E., & Davidovitz, L. (2005). Online versus traditional teaching evaluation: Mode can matter. *Assessment and Evaluation in Higher Education, 30* (6), 581-592.

Gosling, S. D., Vazire, S., Srivastava, S. J., & Oliver P. (2004). Should we trust web-based studies?: A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist, 59*(2), 93-104.

Hakstian, R.A., & Whalen, T.E. (1976), A k-sample significance test for independent alpha coefficients, *Psychometrika, 41*, 219-231.

Kleiman, T., & Gati, I. (2004). Challenges of internet-based assessment: Measuring career decision-making difficulties. *Measurement and Evaluation in Counseling and Development, 37*, 41-55.

McCue, P., Martin, C. R., Buchanan, T., Rodgers, J., & Scholey, A. B. (2003). An investigation into the psychometric properties of the Hospital Anxiety and Depression Scale in individuals with chronic fatigue syndrome. *Psychology, Health & Medicine, 8*(4), 425-439.

McKelvie, S. J. (2004). Is the neuroticism scale of the Eysenck Personality Inventory contaminated by response bias? *Personality and Individual Differences, 36*(4), 743-755.

Mertler, C. A. & Earley, M. A. (2002, October). *The mouse or the pencil? A psychometric comparison of Web-based and traditional survey methodology*. Paper presented at the annual meeting of the Mid-Western Educational Research Association, Columbus, Ohio.

Mertler, C. A., & Earley, M. A. (2003, April). *A Comparison of the Psychometric Qualities of Surveys Administered by Web and Traditional Methods*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Mertler, C. A. (2003). Patterns of response and nonresponse from teachers to traditional and web surveys. *Practical Assessment, Research & Evaluation*, 8(22). Retrieved July, 27, 2010 from http://PAREonline.net/getvn.asp?v=8&n=22

Miller, E. T., Neal, D. J., Roberts, L. J., Baer, J. S., Cressler, S. O., Metrik, J., & Marlatt, G. A. (2002). Test–retest reliability of alcohol measures: Is there a difference between internet-based assessment and traditional methods? *Psychology of Addictive Behaviors, 16*(1), 56-63.

Osborne, J. W., & Blanchard, M. R. (2011). Random responding from participants is a threat to the validity of social science research results. *Frontiers in Psychology, 1*, 1-7.

Paulhus, D. L. (1991). Measurement and control of response bias. In: J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of Personality and Social Psychological Attitudes* (pp. 17-59). San Diego, CA: Academic Press.

Potosky, D., & Bobko, P. (1997). Computer versus paper-and-pencil administration mode and response distortion in noncognitive selection tests. *Journal of Applied Psychology, 82*(2), 293-299.

Riva, G., Teruzzi, T., & Anolli, L. (2003). The use of the Internet in psychological research: Comparison of online and offline questionnaires. *Cyber Psychology & Behavior, 6*(1), 73-80.

Rodriguez, M. C., & Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychological Methods, 11*(3), 306-322.

Ruble, T. L., & Stout, D. E. (1990). Reliability, construct validity, and response-set bias of the revised Learning-Style Inventory (LSI-1985). *Educational and Psychological Measurement, 50*(3), 619-629.

Ruble, T. L., & Stout, D. E. (1991). Reliability, classification stability, and response-set bias of alternate forms of the Learning-Style Inventory (LSI-1985). *Educational and Psychological Measurement, 51*(2), 481-489.

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychological Methods, 8*(2), 206-224.

Spector, P. E., Van Katwyk, P., Brannic, M. T., & Chen, P. Y. (1997). When two factors don't reflect two constructs: How item characteristics can produce artifactual factors. *Journal of Management, 23*(5), 659-677.

Sullivan, B. F., & Scandell, D. J. (2003). Psychological needs and response bias: An examination of Paulhus and John's reformulation. *North American Journal of Psychology, 5*(2), 279-287.

Tibbles, A. C., Waalen, J. K. & Hains, F. (1998). Response set bias, internal consistency and construct validity of the Oswestry Low Back Pain Disability Questionnaire. *Journal of Canadian Chiropractic Association,* 42(3), 141-149.

Whittier, D. K., Seeley, S., & St. Lawrence, J. S. (2004). A comparison of web- with paper-based surveys of gay and bisexual men who vacationed in a gay resort community. *AIDS Education and Prevention, 16*(5), 476-485.

Citation:

Peer, Eyal & Gamliel, Eyal (2011) Too reliable to be true? Response bias as a potential source of inflation in paper-and-pencil questionnaire reliability. *Practical Assessment, Research & Evaluation*, 16(9). Available online: http://pareonline.net/getvn.asp?v=16&n=9.

Acknowledgement

We wish to thank Meni Berger, of SPSS Inc. Israel, for his assistance in some of the data analyses.

Authors:

Eyal Peer
School of Education
Hebrew University of Jerusalem
Israel.
eyal.peer [at] mail.huji.ac.il

Eyal Gamliel
Behavioral Sciences Department
Ruppin Academic Center
Israel
eyalg [at] ruppin.ac.il