Short Communication

Using pre- and post-survey instruments in interventions: determining the random response benchmark and its implications for measuring effectiveness

George C Davis[1,2,*], Ranju Baral[3], Thomas Strayer[4] and Elena L Serrano[1]

[1]Department of Human Nutrition, Foods, and Exercise, Virginia Tech University, Blacksburg, VA, USA: [2]Department of Agricultural and Applied Economics, Virginia Tech University, 214 Hutcheson Hall, Blacksburg, VA 24061, USA: [3]Global Health Group, University of California San Francisco, Global Health Sciences, San Francisco, CA, USA: [4]Translational Biology, Medicine, and Health Program, Virginia Tech University, Roanoke, VA, USA

Submitted 7 October 2016: Final revision received 11 August 2017: Accepted 1 November 2017: First published online 21 December 2017

Abstract

Objective: The present communication demonstrates that even if individuals are answering a pre/post survey at random, the percentage of individuals showing improvement from the pre- to the post-survey can be surprisingly high. Some simple formulas and tables are presented that will allow analysts to quickly determine the expected percentage of individuals showing improvement if participants just answered the survey at random. This benchmark percentage, in turn, defines the appropriate null hypothesis for testing if the actual percentage observed is greater than the expected random answering percentage.

Design: The analysis is demonstrated by testing if actual improvement in a component of the US Department of Agriculture's (USDA) Expanded Food and Nutrition Education Program is significantly different from random answering improvement.

Setting: USA.

Subjects: From 2011 to 2014, 364 320 adults completed a standardized pre- and post-survey administered by the USDA.

Results: For each year, the statement that the actual number of improvements is less than the expected number if the questions were just answered at random cannot be rejected. This does not mean that the pre-/post-test survey instrument is flawed, only that the data are being inappropriately evaluated.

Conclusions: Knowing the percentage of individuals showing improvement on a pre/post survey instrument when questions are randomly answered is an important benchmark number to determine in order to draw valid inferences about nutrition interventions. The results presented here should help analysts in determining this benchmark number for some common survey structures and avoid drawing faulty inferences about the effectiveness of an intervention.

Keywords
Pre- and post-survey
Evaluation
Random answering
Inference bias
Baseline benchmark

In efforts to measure effectiveness, pre- and post-surveys are common in nutrition interventions, given their simplicity, low response burden and ease of administration[1–6]. Prior to the intervention a pre-survey is administered, and the same survey is administered again after the intervention. The survey usually consists of multiple questions with either dichotomous (e.g. yes/no) or polychotomous (e.g. Likert-scale responses: 0 = very low, 1 = low, 2 = medium, 3 = high, 4 = very high) responses. A 'positive' change from the pre- to the post-survey is then considered as demonstrating the effectiveness of the intervention or simply as improvement ('positive' includes any required reverse coding). This improvement may be reported in various forms (e.g. average score change, number of questions improved on, percentage of individuals showing improvement). The focus here is on the percentage of individuals showing improvement from the pre- to the post-survey, as the main intervention intent is to impact individuals not scores (i.e. an improved average score tells nothing about the number of individuals improving).

[*] *Corresponding author:* Email georgedavis@vt.edu

https://doi.org/10.1017/S1368980017003639 Published online by Cambridge University Press

To draw valid inference about the effect of an intervention, it is important to know the expected results if the questions were simply answered at random (e.g. simply guessing in an objective, right or wrong type question). In statistics, the random effect forms the basis for determining the appropriate null hypothesis, the appropriate test, and therefore determining the appropriate (valid) conclusion to draw about the effectiveness of the intervention. In a pre- and post-survey, if the intervention is effective, the answers should show a pattern different from being randomly answered.

The purpose of the present short communication is twofold. First, the communication shows that random answering in a typical pre/post format can lead to a surprisingly large percentage of individuals 'showing improvement', which can give the misleading impression that the intervention is effective. Second, and more constructively, the steps, formulas and a table are provided to aid analysts in quickly determining the expected percentage of respondents showing improvement if the questions were simply answered at random. This number can then be used, as the null hypothesis, in testing if the observed percentage is significantly different from the random answering percentage. We provide an illustrative example using data from an annual nationwide pre/post survey conducted by the US Department of Agriculture (USDA) in its Expanded Food and Nutrition Education Program (EFNEP).

Methods

A simple example gives the basic intuition before turning to the general case.

A simple example

Suppose a nutrition intervention is designed to improve fresh fruit intake. The pre/post survey has one question: 'On a daily basis, how frequently do you eat fresh fruit?' The possible response answers are: 1 = never, 2 = seldom, 3 = sometimes, 4 = often and 5 = always. Table 1 shows all possible answers from the pre- and post-survey (the event space). The rows represent the five possible responses to the pre-survey and the columns the five possible responses to the post-survey. There is a total of twenty-five (= 5 × 5) possible answer combinations. An improvement on the question is defined as a higher response on the post-survey than the pre-survey, so there are ten possible improvement answer combinations, which are shown in the shaded upper off-diagonal cells. Random answering would imply an equal probability for any cell, so the probability of showing an improvement on the question is 10/25 or 0·40. Suppose 100 people participated in the intervention. If individuals are (independently) answering at random, the expected number of individuals showing improvement is then $100 \times 0{\cdot}40 = 40$, or 40 % are expected to show improvement just by chance. This establishes a well-defined quantitative benchmark for analysis and testing. Without this benchmark one does not know the relevant comparison for statistical testing and drawing correct conclusions about the intervention. And, regardless of statistical significance, in many interventions a 40 % improvement rate would be considered clinically significant when in fact this is the expected random answering percentage. This random response information is normally not provided in pre- and post-survey based studies, but is very useful for benchmarking effects and drawing valid inferences.

Table 1 All possible answer combinations for a pre- and post-survey question with a five-point scale. The shaded area shows improvement events

Pre/Post	1	2	3	4	5
1	11	12	13	14	15
2	21	22	23	24	25
3	31	32	33	34	35
4	41	42	43	44	45
5	51	52	53	54	55

The general approach in two steps

Most pre/post surveys consist of multiple questions and in this case the analyst is likely interested in several alternative probabilities. Here we focus on two. First, out of n questions, what is the probability of showing an improvement in all n questions if the questions are answered at random? Second, what is the probability of improving on at least one question, at least two questions, etc., if the questions are answered at random? The answers to these questions are related and involve two steps. First, the survey response structure can be used to determine the probability of showing an improvement in each question, call it P. Second, this probability from step one can be used in the binomial probability distribution to determine the relevant probabilities for the number of questions of interest.

Step one

Generalizing the simple example, suppose every question has a Likert-scale response consisting of k possible answers. The total possible answer combinations from the pre- and post-survey will then be k^2 (the event space). The number of possible improvements in the Likert scale from the pre- to the post-survey is then $(k^2 - k) \div 2$. Random answering implies that the probability of observing an improvement in a question is the number of possible improvements divided by the entire possible event space or $P = (k^2 - k) \div 2k^2 = (k - 1) \div 2k$. All of this is just the generalization of the simple example above where $k = 5$.

Step two

Under random answering, along with the probability of an improvement on any one question in the survey from step one, the binomial distribution gives the probability of improvement (define as a success) on any number of questions[7] as:

$$P(y)=\frac{n!}{y!\,(n-y)!}P^{y}(1-P)^{n-y}, \qquad (1)$$

where n is the number of questions in the survey, y is the number of successes (number of questions improved on) and P is the probability of improving on a single question at random from step one. Clearly the probability of y depends on the number of responses k through P, but also on the number of questions n and the number of improvement responses y considered to be the relevant threshold.

While the formulas above could be used for any number of questions and responses, Table 2 gives the results for some typical survey structures. The three subsections correspond to a two-point scale ($k=2$: e.g. true/false, yes/no), a three-point scale ($k=3$: e.g. never, sometimes and always) and a five-point scale ($k=5$: e.g. never, seldom, sometimes, often and always). For each subsection, the n rows refer to the number of questions in the survey. The column labelled 'All' ($y=n$) gives the probability of showing an improvement in all n questions. That is, 'All' means improvement on every question. The other columns show the probabilities of improving on 'at least' the number of questions given by the inequality (e.g. at least one, any one, $y\geq 1$) and these come from using equation (1) with some basic properties of probabilities[7]. For example, consider the fifth row of the top subsection of Table 2. This corresponds to a survey with $n=5$ questions and each question has two possible answers ($k=2$). The probability of improving on all questions ($y=5$) by answering at random is 0·00. However, note the probability of improving on 'at least' one question by answering at random is very high at 0·76. This implies that if improving on at least one question (or equivalently, more than zero) is the criterion for measuring improvement and 100 people did the pre/post survey, then we would expect seventy-six out of 100 people to show improvement on one or more questions, even if they were just answering the five questions at random. Note, after determining the value of P for the survey structure from step one, equation (1) can be used to determine the probability of answering any subset y out of n questions correctly.

There are some important general patterns to observe in Table 2, especially with respect to the 'at least' columns. For any fixed number of responses (i.e. a given value of k), all the 'at least' probabilities increase as the number of questions n asked increases (i.e. within any subsection the probabilities increase as you go down the rows). Stated more simply, just adding more questions to a survey will increase the probability of showing an improvement. Also note for a given k and n, all the 'at least' probabilities increase as the 'at least' threshold decreases (i.e. within any subsection the probabilities increase as you go across columns from right to left). So, decreasing the improvement threshold, from say $y=4$ to 3 to 2 to 1, will increase the probability of showing improvement. Finally, for any given number of questions n, looking across the different values of k reveals that the probabilities increase as the number of possible responses increases from $k=2$ to $k=3$ to $k=5$. So just increasing the number of response categories increases the probability of showing an

Table 2 Probabilities of random answering for various survey structures and improvement criterion

		Number of questions showing improvement (y)							
		All	At least						
Answer responses (k)	Questions (n)	$y=n$	$y\geq 1$	$y\geq 2$	$y\geq 3$	$y\geq 4$	$y\geq 5$	$y\geq 6$	$y\geq 11$
$k=2$	$n=1$	0·25	0·25						
	$n=2$	0·06	0·44	0·06					
	$n=3$	0·02	0·58	0·16	0·02				
	$n=4$	0·00	0·68	0·26	0·05	0·00			
	$n=5$	0·00	0·76	0·37	0·10	0·02	0·00		
	$n=10$	0·00	0·94	0·76	0·47	0·22	0·08	0·02	
	$n=20$	0·00	1·00	0·98	0·91	0·77	0·59	0·38	0·00
$k=3$	$n=1$	0·33	0·33						
	$n=2$	0·11	0·56	0·11					
	$n=3$	0·04	0·70	0·26	0·04				
	$n=4$	0·01	0·80	0·41	0·11	0·01			
	$n=5$	0·00	0·87	0·54	0·21	0·05	0·00		
	$n=10$	0·00	0·98	0·90	0·70	0·44	0·21	0·08	
	$n=20$	0·00	1·00	1·00	0·98	0·94	0·85	0·70	0·04
$k=5$	$n=1$	0·40	0·40						
	$n=2$	0·16	0·64	0·16					
	$n=3$	0·06	0·78	0·35	0·06				
	$n=4$	0·03	0·87	0·52	0·18	0·03			
	$n=5$	0·01	0·92	0·66	0·32	0·09	0·01		
	$n=10$	0·00	0·99	0·95	0·83	0·62	0·37	0·17	
	$n=20$	0·00	1·00	1·00	1·00	0·98	0·95	0·87	0·13

improvement as well. In summary, the general result is that a pre/post survey format with many questions, with many response categories and a low improvement threshold is more likely to show improvement simply by chance.

An application and test

To demonstrate the usefulness of these results, we analyse some publicly available data from the USDA related to the EFNEP. The EFNEP is one of the largest nutrition education programmes in the USA as it is administered in all fifty states every year and involves an education curriculum[8]. A standardized ten-question pre- and post-survey is administered to all adult participants. Each state enters its individual-level data into the national Nutrition Education Evaluation and Reporting System and USDA then aggregates the data and reports nationwide 'impact' indicators, which are simply the number of participants that improved on the standardized pre- to post-survey. These data are reported every year and have even been used recently to look at the cost-effectiveness of the EFNEP[6]. The ten survey questions are designed to cover three different domains: food resource management practices (FRMP), nutrition practices (NP) and food safety practices (FSP). For brevity we just focus on the FRMP. The FRMP component contains four questions related to frequency of food management practices with five-point Likert scale responses (1 = do not do, 2 = seldom, 3 = sometimes, 4 = most of the time and 5 = almost always). An individual is considered as showing improvement in FRMP by the USDA if he/she improves on at least one of the four questions. So from Table 2, this implies $k=5$, $n=4$ and the probability of improving on at least one question when randomly answering is 0·87. This in turn implies 87 % of the participants are expected to show improvement simply by answering at random. Without working through the math an 87 % improvement would seem quite impressive, when in fact it is what is expected with random answering.

With the expected proportion value in hand under the null hypothesis of random answering, testing the statistical significance can be done with a proportions test[7]. The null and alternative hypotheses, along with the test statistic and rejection region, are as follows:

$$H_0:\hat{\pi}\leq\pi_0 \quad H_a:\hat{\pi}>\pi_0$$

$$z=\frac{\hat{\pi}-\pi_0}{\sqrt{N^{-1}\pi_0(1-\pi_0)}} \quad \text{Reject } H_0 \text{ if } z>z_\alpha$$

where $\hat{\pi}$ is the observed proportion of participants showing improvement, π_0 is the expected proportion if questions are answered at random (e.g. 0·87), N is the number of participants completing both pre- and post-surveys, and α is the chosen significance level.

Using data found in the USDA impact reports, the null hypothesis that the actual percentage is less than the expected percentage under random answering is tested for the FRMP 2010–2014[9]. Table 3 gives the results. The N row gives the number of individuals completing the pre- and post-survey in each year. The number of individuals ranges from about 68 000 (2014) to 76 000 (2010) and the actual observed proportion that showed improvement in one or more questions was about 84 %, which sounds impressive. However, as demonstrated above, if individuals just answered at random, the expected proportion would be 87 %. Using the above z-test statistic, the P values for the test statistics indicate for all years we cannot reject the null hypothesis that the actual proportion is less than the proportion we would expect if the questions were answered at random. Simply stated, the actual proportion showing improvement is less than what we would expect if they were all just answering the questions randomly.

Discussion

This short communication is a cautionary note on utilizing the information collected from certain pre- and post-survey instruments. The probability of showing an improvement can be surprisingly high even if the questions are answered at random. This probability is normally not reported in pre-/post-survey analyses. To assist analysts, the steps, formulas and a table for determining the probability of showing an improvement by random answering are provided, which should prove useful when designing a

Table 3 US Department of Agriculture Expanded Food and Nutrition Education Program food resource management practices: expected *v.* actual improvement proportions and test results, 2010–2014

	2010	2011	2012	2013	2014
N	76 071	75 418	73 958	71 014	67 859
Expected improvement proportion	0·87	0·87	0·87	0·87	0·87
Actual improvement proportion	0·84	0·83	0·85	0·84	0·84
z-Statistic	–26·50	–29·92	–20·33	–21·97	–22·18
P value	0·99	0·99	0·99	0·99	0·99
Decision	No reject	No reject	No reject	No reject	No reject

N is the number of individuals completing both the pre- and post-survey. *P* value is for the null hypothesis that the actual improvement proportion is less than the expected improvement proportion when questions are answered at random.

https://doi.org/10.1017/S1368980017003639 Published online by Cambridge University Press

pre- and post-survey instrument and using the instrument to evaluate an intervention.

Pre- and post-surveys have a long history in evaluating interventions, especially nutrition interventions. As with all instruments they have pros and cons. What does this research imply? It does not imply that pre- and post-surveys are flawed and uninformative. Some research indicates this type of low-response-burden survey may be valid and reliable in correlating with more time-intensive, accurate assessment metrics in some applications but not in others and this is an important ongoing research area(10–12). Our concern here is not with this correlation validity, but simply how the data from such surveys are presented and analysed. Consequently, in our application, our analysis does not mean the EFNEP is ineffective or effective. To draw this conclusion is to miss the main point of the communication. One would not claim an intervention to increase fruit consumption was ineffective because a 24 h dietary recall did not show any change in energy intake. The problem would not be the intervention or 24 h dietary recall, the problem is the analyst is utilizing the data from the 24 h dietary incorrectly to answer the question of interest. The logic here is similar. The problem is the improvement metric, not the instrument or programme. Exceeding a very low threshold for showing improvement is unlikely to reveal anything meaningful about the effectiveness of a programme.

The implication of this research is that the analyst should think carefully about what random responding, the appropriate null, would imply for the proposed measure and at a minimum test against that null or, better yet, use a more sophisticated measure that would not be subject to the problem explained here. There is a continuum of more sophisticated and reliable techniques an analyst could pursue to improve the analysis and still use the data from the pre- and post-survey. Given that many of the pre- and post-surveys are used in a nutrition education context, the logical place to look for more sophisticated methods is in the general education literature. The most common approach found in that literature for measuring effectiveness via a testing instrument is some type of Rasch model(13). Regardless, knowing the expected responses under random answering is an important benchmark to report and consider.

Acknowledgements

Financial support: Partial funding for this project was provided by the USDA, National Institute of Food and Agriculture, National Research Initiative, Human Nutrition and Obesity, Project 2009-55215-05074. The USDA played no role in the study design, analysis, interpretation of the results, or preparation of the manuscript. *Conflict of interest:* There are no conflicts of interest. *Authorship:* G.C.D. and R.B. were responsible for the manuscript idea, development and implementation of the initial analysis. T.S. provided assistance with subsequent analysis and E.L.S. provided interpretative insights on analysing the data and results. G.C.D. wrote the initial draft of the manuscript and all other authors assisted in revisions. *Ethics of human subject participation:* There are no ethical considerations associated with the manuscript. Review by the institutional review board was not required for this study because human subjects were not involved, as per US Department of Health and Human Services guidelines.

References

1. Andreyeva T, Middleton AE, Long MW *et al.* (2011) Food retailer practices, attitudes, and beliefs about the supply of healthy foods. *Public Health Nutr* **14**, 1024–1031.
2. Song HJ, Gittelsohn J, Kim M *et al.* (2009) A corner store intervention in a low-income urban community is associated with increased availability and sales of some healthy foods. *Public Health Nutr* **12**, 2060–2067.
3. Martínez-Donate AP, Riggall AJ, Meinen AM *et al.* (2015) Evaluation of a pilot healthy eating intervention in restaurants and food stores of a rural community: a randomized community trial. *BMC Public Health* **15**, 136.
4. Escaron AL, Martinez-Donate AP, Riggall AJ *et al.* (2016) Developing and implementing 'Waupaca Eating Smart': a restaurant and supermarket intervention to promote healthy eating through changes in the food environment. *Health Promot Pract* **17**, 265–277.
5. Lee RM, Rothstein JD, Gergen J *et al.* (2015) Process evaluation of a comprehensive supermarket intervention in a low-income Baltimore community. *Health Promot Pract* **16**, 849–858.
6. Baral R, Davis GC, Blake S *et al.* (2013) Using national data to estimate average cost effectiveness of EFNEP outcomes by state/territory. *J Nutr Educ Behav* **45**, 183–187.
7. Ott RL & Longnecker M (2001) *An Introduction to Statistical Methods and Analysis*, 5th ed. Pacific Grove, CA: Duxbury.
8. US Department of Agriculture (2016) Expanded Food and Nutrition Education Program (EFNEP). http://www.nifa.usda.gov/program/expanded-food-and-nutrition-education-program-efnep (accessed September 2016).
9. US Department of Agriculture (2016) Expanded Food and Nutrition Education Program. National Data. Years 2010–2014. http://www.nifa.usda.gov/efnep-national-data-reports (accessed September 2016).
10. Murphy SP, Kaiser LL, Townsend MS *et al.* (2001) Evaluation of validity of items for a food behavior checklist. *J Am Diet Assoc* **101**, 751–761.
11. George GC, Milani TJ, Hanss-Nuss H *et al.* (2004) Development and validation of a semi-quantitative food frequency questionnaire for young adult women in the southwestern United States. *Nutr Res* **24**, 29–43.
12. Lim SS, Gold A, Gaillard PR *et al.* (2015) Validation of 2 brief fruit and vegetable assessment instruments among third-grade students. *J Nutr Educ Behav* **47**, 446–451.
13. Bond TG & Fox CM (2007) *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, 2nd ed. Mahwah, NJ: LEA Publishers.

https://doi.org/10.1017/S1368980017003639 Published online by Cambridge University Press