Available online at www.sciencedirect.com

ScienceDirect

Health Professions Education 5 (2019) 76–81

www.elsevier.com/locate/hpe

When Negative Turns Positive and Vice Versa: The Case of Repeated Measurements

Jimmie Leppink[1]

School of Health Professions Education, Maastricht University, PO Box 616, 6200 MD Maastricht, The Netherlands

Received 28 March 2017; accepted 28 March 2017
Available online 23 April 2017

Abstract

Background: In increasing numbers of studies, participants are measured on the same variables of interest – such as task performance and effort associated with task performance – several times. Three statistical approaches that are encountered in health professions education research in this context are: aggregation, multiplication, and repeated measures analysis. In the *aggregation* approach, repeated measurements are reduced to average scores (i.e., aggregates) for each measure and these are used in subsequent analysis. In the *multiplication* approach, repeated measurements are treated as if they came from different instead of from the same respondents. In the *repeated measures* approach, the repeated measurements are treated as is. While the aggregation and multiplication approach are frequently encountered, the repeated measures approach is not.
Method: Through a simulated data example that incorporates features from studies on this kind of data encountered in the literature, this article compares the three aforementioned approaches in terms of information and statistical validity.
Results: The comparison illustrates that, contrary to repeated measures analysis, the aggregation and multiplication approach fail to capture essential information from repeated measurements data and can result in erroneous implications for future research and practice (i.e., ecological fallacy).
Discussion: The findings from this comparison have implications for all statistical techniques that are based on correlation and regression; failing to account for repeated measurements structures in data can distort correlations and all statistics based on them.

Keywords: Repeated measurements; Ecological fallacy; Repeated measures analysis

1. Introduction

In increasing numbers of studies, participants are measured on the same variables of interest two or more times. For instance, researchers may ask participants to rate the mental effort invested in task performance[1] a number of times within a given time interval.[2,3] Other researchers may collect performance data on a series of tasks, for example interpreting eight electrocardiograms (ECGs).[4] In some cases, researchers collect such repeated measurements on more than one variable simultaneously, for instance both learning task performance and mental effort[5] or both interpretation time and interpretation performance.[4]

E-mail address: jimmie.leppink@maastrichtuniversity.nl

[1]He is currently Assistant Professor in Education at the School of Health Professions Education, Maastricht University, The Netherlands.

Peer review under responsibility of AMEEMR: the Association for Medical Education in the Eastern Mediterranean Region

http://dx.doi.org/10.1016/j.hpe.2017.03.004

1.1. Statistical approaches to repeated measurements

Three statistical approaches that are encountered in the context of repeated measurements in health professions education research are: aggregation, multiplication, and repeated measures analysis .[6]

In the *aggregation* approach, repeated measurements are aggregated to single average (or sum) scores and these aggregates are then used in subsequent analysis. For example, recent work has suggested that the average of a series of mental effort ratings during an activity tends to be lower than a single mental effort rating at the end of that activity and that it is therefore better to use the average of repeated measurements in analysis instead of a single effort rating.[2,3] While this is a very important finding for a variety of reasons, through aggregation of repeated measurements to a single average score, data from N number of participants who were measured on the same variable – for instance performance on each of a series of learning tasks (e.g., correct/incorrect) and/or self-reported mental effort (e.g., 1–9) after each learning task[3] – k times are reduced to N number of independent observations. As such, the aggregation approach treats repeated measurements from the same participants as 'all being the same' or perfectly related. Although this approach is very common in education research, we lose all within-participant variation and all potentially useful information about for instance group-by-time interaction effects with it.[7]

In the *multiplication* approach, data from N number of participants who were measured on the same variable k times are treated as if they come from N x k different participants. For example, performance (e.g., correct/incorrect) and self-reported mental effort (e.g., 1–9) scores on each of eight ($k=8$) ECGs obtained from each of six residents ($N = 6$) are then treated as $8 \times 6 = 48$ independent observations, and a correlation is calculated between the 48 performance scores and the 48 mental effort scores[4] as if 48 residents ($N = 48$) each performed and self-reported their mental effort on a single ($k = 1$) ECG. This approach is sometimes justified by pointing at a small sample size due to not being able to recruit more participants, due to dropout, or due to a mixed-methods character of a study where for instance interviews have been conducted and quantitative data may be used to triangulate some of the interview findings or vice versa. By multiplying the number of participants (N) and the number of measurements per participant (k), we assume no relation whatsoever between repeated measurements from the same participants.

In the *repeated measures approach*, repeated measurements data are actually treated as such. The number of independent observations in this approach typically lies somewhere between that of the aggregation approach (N) and that of the multiplication approach ($N \times k$); the more correspondence between repeated measurements from the same participants, the more the number goes down towards N. Doing so, the repeated measures approach recognizes that neither extreme (perfect relation and no relation, respectively) is realistic but that some dependency exists: repeated measurements from the same participants tend to be more similar than measurements from different participants.[6] Moreover, contrary to the other two approaches, the within-participant variation and information about group-by-time interaction effects are preserved in the repeated measures approach. For example, consider that two groups of fifty residents ($N = 100$, two times $n = 50$) complete a series of five practice tasks ($k = 5$) and rate their mental effort after each task. If the groups do not differ in effort rating on the first two or three tasks but differ substantially on the last two tasks, there is a group-by-time interaction effect. In the aggregation approach, researchers would calculate the mental effort averaged across five tasks for each of $N = 100$ participants, do a t-test on the difference between the two groups (of $n = 50$ each), and fail to see the interaction.[7] In the multiplication approach, researchers would do a t-test on $N \times k = 100 \times 5 = 500$ ratings on the difference between the two 'groups' (of $n \times k = 50 \times 5 = 250$ each); they might find a statistically significant difference between 'groups' through a gross exaggeration of the sample size[6] but would fail to see the interaction. In the repeated measures approach, the true number of observations lies somewhere between 100 and 500 (i.e., somewhere between 50 and 250 per group) and the interaction effect can be tested directly.[7]

1.2. The current study: the relation between variables that are measured repeatedly

While the aggregation and multiplication approach are frequently encountered in situations where researchers have repeated measurements data, the repeated measures approach is encountered less frequently.[7] In this study, the three approaches are compared in terms of information and statistical validity using a simulated data example that incorporates features from studies on this kind of data encountered in the literature (e.g., in the context of residents interpreting ECGs).[4] The main question we

want to address with each of the three approaches is whether effort ratings can provide a linear prediction of performance scores. More specifically, does a resident's performance tend to increase or decrease with increases in effort?

2. Method

Since small sample studies are not uncommon in health professions education research and a small sample example makes it easy to illustrate differences between statistical approaches when used on the same dataset, data were simulated for a small sample study.

2.1. Participants

In this study, eight residents ($N = 8$) each performed six practice tasks ($k = 6$).

2.2. Measures

Each practice task yielded a self-reported effort rating on a visual analog scale from 0 (minimum) to 100 (maximum) and a performance score ranging from 0 (minimum) to 100 (maximum).

2.3. Procedure

Residents individually performed the first practice task and, immediately after completion, rated their effort invested in performing that task. This process was repeated for each resident until the eighth task had been completed. Since residents did not receive performance feedback at any time during the study, effort ratings were not influenced by any form of performance feedback.

2.4. Analysis

In the aggregation approach, we aggregate the six performance scores obtained for each resident to a single average performance score (0–100) per resident and we aggregate the six effort scores obtained for each resident to a single average effort score (0–100) per resident. Hence, since we have $N = 8$ residents, we obtain 8 average performance scores and 8 average effort scores. Next, we perform linear regression using average performance as response variable and average effort as predictor variable.

In the multiplication approach, we reason that we actually do not have such a small sample: since each of eight residents is measured six times on both performance and effort, we have a total of 48 performance scores and 48 effort scores. In other words, we treat the 48 scores as if 48 residents each performed one task and provided one effort rating. We perform linear regression using performance as response variable and effort as predictor variable.

Table 1
Data of this example: performance (0–100) and effort (0–100) per resident per task.

Resident	Performance (task 1, 2, 3, 4, 5, 6)	Effort (task 1, 2, 3, 4, 5, 6)
A	81, 82, 84, 85, 84, 87	10, 17, 20, 22, 24, 26
B	74, 77, 79, 81, 82, 85	13, 18, 21, 23, 25, 30
C	73, 74, 73, 77, 78, 80	16, 17, 19, 21, 23, 27
D	66, 70, 69, 72, 73, 76	21, 23, 24, 27, 29, 33
E	62, 64, 66, 68, 70, 73	23, 25, 28, 29, 31, 34
F	58, 60, 64, 65, 68, 70	27, 29, 33, 34, 38, 41
G	54, 55, 57, 59, 62, 63	30, 32, 37, 41, 44, 46
H	50, 53, 54, 55, 61, 62	31, 34, 35, 36, 39, 40

In the repeated measures approach, we recognize that we do have 8 residents that altogether provide 48 scores for each of performance and effort but the effective sample size lies somewhere between 8 (perfect correlation between repeated measurements) and 48 (no correlation between repeated measurements) because the fact that each resident was measured six times likely induces a so-called within-resident between-measurements correlation.[6] We therefore perform two-level regression analysis with performance as response variable and effort as predictor in which resident and measurement occasion are treated as hierarchical levels.[7] The latter allows us to include resident-level random effects, 'random' because we typically conveniently assume random sampling of residents and wish to generalize the findings from the

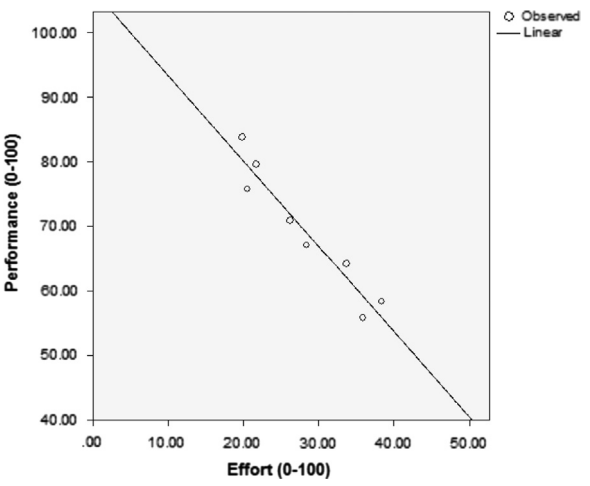

Fig. 1. Scatterplot resulting from the aggregation approach.

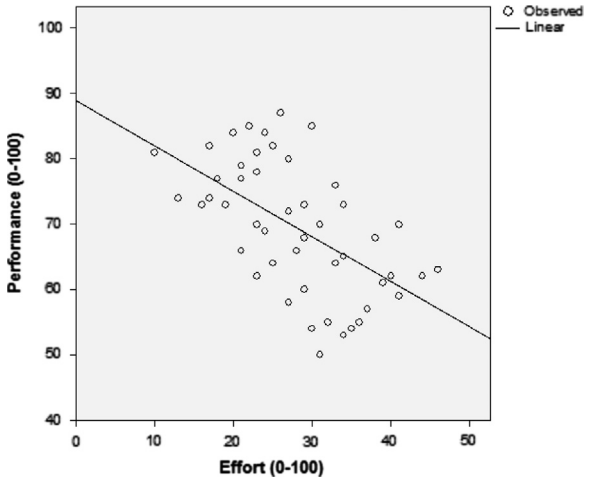

Fig. 2. Scatterplot resulting from the multiplication approach.

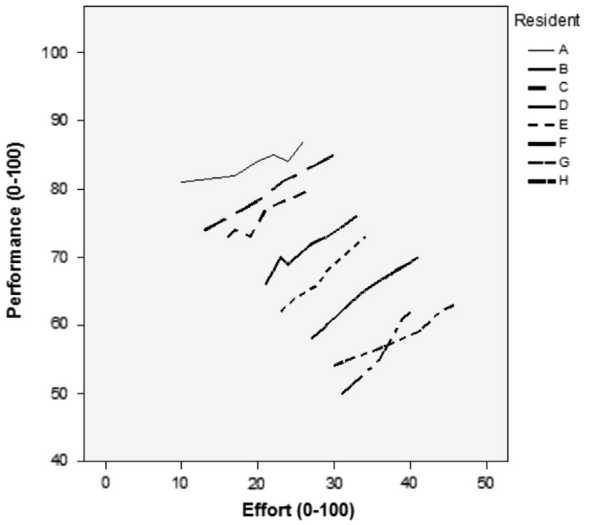

Fig. 3. Plot resulting from the repeated measures approach.

residents in our study to a population these residents were sampled from.[6] Including these random effects enables researchers to estimate the correlation between measurements and account for it while analyzing relations of interest.[7–10] In this case, the relation of interest is that between performance and effort. To keep the example simple, we include resident-level random intercepts, that is: we assume that different residents tend to differ in performance and effort and that creates a correlation between measurements that is proportional with these differences .[9]

3. Results

Table 1 presents the data.

In this section, the findings with regard to the relation between performance and effort are presented for each of the three approaches.

3.1. Approach A: aggregation

Fig. 1 presents the scatterplot resulting from the aggregation approach.

The combinations of performance and effort score of the eight residents are situated such that a *negative* linear relation summarizes the findings best. The linear regression equation is:

$$\textit{Expected performance score} = 106.571 - (1.323 \times \textit{Effort score}).$$

The regression line is the best fitting straight line through the cloud of 8 dots and indicates that, compared with others, a resident who has to invest an above average effort tends to perform below average. Note that this does not respond to our question whether a resident's performance tends to increase or decrease with increases effort. The latter is a question that we cannot answer with the aggregation approach because we lose all within-resident variation when we aggregate. Stating that Fig. 1 and the resulting regression equation address our question of interest would be like saying that a given resident tends to turn into another resident once investing more effort, and that is of course a thought that scientifically does not make a lot of sense.

3.2. Approach B: multiplication

Fig. 2 presents the scatterplot resulting from the multiplication approach.

Although, in line with the aggregation approach, a *negative* relation summarizes the findings best, the resulting linear regression equation is:

$$\textit{Expected performance score} = 88.907 - (0.693 \times \textit{Effort score}).$$

The regression line is the best fitting straight line through the cloud of 48 dots. One could interpret this line as follows: resident-by-task combinations that yield higher effort ratings tend to yield a higher performance. Again, we cannot translate this into anything of meaning to our question whether a resident's performance tends to increase or decrease with increases in effort, this time because we fail to distinguish within-resident variation between-resident variation; they are thrown on one pile.

3.3. Approach C: repeated measures

Fig. 3 presents the plot resulting from the repeated measures (i.e., two-level regression) approach.

In contrast to the other two approaches, the repeated measures approach indicates that a *positive* linear relation summarizes the findings best, with the resulting linear regression equation being:

$$Expected\ performance\ score = 50.234 + (0.686 \times Effort\ score).$$

In this two-level regression model, the regression line for the expected performance score is the average regression line of the 8 individual regression lines, and the individual regression lines clearly demonstrate that when a person does more effort the performance tends to be better as well. Hence, the repeated measures approach provides us with an answer to our question.

4. Discussion

The comparison in this article illustrates that, contrary to repeated measures analysis, the aggregation and multiplication approach fail to capture essential information from repeated measurements data and can result in an *ecological fallacy*: aggregation (in the aggregation approach) or disaggregation (in the multiplication approach) of repeated measurements data provides a substantially different picture of a phenomenon of interest compared to a model that appropriately accounts for repeated measurements structures in the data.[7–9] In the aggregation approach all within-resident variation is lost, and in the multiplication approach all within-resident and between-resident variation are thrown together; repeated measures analysis distinguishes between these two sources of variance, yields a regression line that is the average of individual regression lines, and as such provides an answer to our question whether a resident's performance tends to increase or decrease with increases in effort (i.e., in this example, it tends to increase).

4.1. Take home message: treat repeated measurements data as is

The findings from the comparison in this article have implications for all statistical techniques that are based on correlation or regression; failing to account for repeated measurements structures in data can distort correlations and all statistics based on them. In randomized controlled experiments or quasi-experimental studies in which the interest lies in differences between groups that have been measured several times, the aggregation approach results in a total loss of information with regard to group-by-time interaction[7] – which is often of primary interest especially in experiments – and the multiplication approach tends to elevate Type I error (i.e., false positive) probability for differences between groups and Type II error (i.e., false negative) probability for differences within groups and group-by-time interaction.[6,9] With regard to correlations between variables that have been measured several times, as this article illustrates, both aggregation and multiplication can result in an ecological fallacy.[7–9]

4.2. On the statistical validity of measurement instruments

An exemplar context in which statistical methods are used that are built on correlations is that of examining the statistical validity of measurement instruments. For instance, a group of researchers has 52 residents complete the same fifteen-item questionnaire 4 times and they want to perform factor analysis on the data to examine the factor structure underlying the items .[11] To do so, they multiply the numbers of 52 and 4 measurements to obtain a 'sample size' of 208 measurements (i.e., multiplication approach). Knowing how much correlations can be distorted when using this approach, the factor structure may look drastically different from what it should be. In cases where, due to small samples, two-level factor analysis with respondents and measurement occasion as hierarchical levels[12] does not work, other alternatives that minimize the risk of ecological fallacy can be found in simultaneous component analysis,[13] principal component analysis per measurement occasion,[11] and multidimensional scaling.[14] In other words, there is no need to resort to aggregation or multiplication of repeated measurements data in the context of research on the statistical validity of measurement instruments and given the risk of ecological fallacy we should consider alternatives just mentioned.

4.3. A cautionary note

This article uses a small sample as example to illustrate the problem of ecological fallacy associated with aggregation and multiplication of repeated measurements data. While samples of this size are not unrealistic in health professions education research, one should carefully reflect on the sample size when intending to do a repeated measures analysis.[7,11] If for logistic or other reasons it is not possible to go

beyond a small sample and two-level regression analysis is not feasible, providing a graphical representation (e.g., Fig. 3) and possibly key descriptive numbers is in any case more meaningful than using more complex statistical tools that make little sense in a given context.

Ethical approval

Not applicable.

Funding

None.

Other disclosures

No conflicts of interest.

References

1. Paas F. Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach. *J Educ Psychol* 1992;84:429–434.
2. Schmeck A, Opfermann M, Van Gog T, Paas F, Leutner D. Measuring cognitive load with subjective rating scales during problem solving: differences between immediate and delayed ratings. *Instr Sci* 2015;43:93–114.
3. Van Gog T, Kirschner P, Kester L, Paas F. Timing and frequency of mental effort measurement: evidence in favour of repeated measures. *Appl Cogn Psychol* 2012;26:833–839.
4. Sibbald M, De Bruin ABH. Feasibility of self-reflection as a tool to balance clinical reasoning strategies. *Adv Health Sci Educ* 2012;17:419–429.
5. Kostons D, Van Gog T, Paas F. Training self-assessment and task-selection skills: a cognitive approach to improving self-regulated learning. *Learn Instr* 2012;22:121–132.
6. Leppink J. Data analysis in medical education research: a multilevel approach. *Perspect Med Educ* 2015;4:14–24.
7. Leppink J, Van Merriënboer JJG. *Educ Technol Soc* 2015;18:230–245.
8. Snijders TAB, Bosker R. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, 2nd ed., London: Sage Publications; 2011.
9. Tan FES. Best practices in analysis of longitudinal data: a multilevel approach. In: Osborne JW, editor. *Best practices in quantitative methods*. London: Sage Publications; 2010. p. 451–470.
10. Verbeke G, Molenberghs G. *Linear mixed models for longitudinal data*. New York: Springer; 2000.
11. Field A. *Discovering Statistics Using IBM SPSS Statistics*, 4th ed., London: Sage Publications; 2013.
12. Byrne BM. *Structural Equation Modeling with Mplus: Basic Concepts, Applications, and Programming*. New York: Taylor & Francis Group; 2012.
13. Ceulemans E, Wilderjans TF, Kiers HAL, Timmerman ME. Multilevel simultaneous component analysis: a computational shortcut and software package. *Behav Res Methods* 2016;48:1008–1020.
14. Borg I, Groenen PJF. *Modern Multidimensional Scaling: Theory and Applications*. New York: Springer; 2005.