Comment on Nalliah and Allareddy (2014): Unwarranted conclusions

Abstract

A PeerJ paper by Nalliah and Allareddy (2014) describes improvement in weaker dental students’ scores (and decline in stronger students’ scores) by use of a unique instructional method. I argue that the causal conclusion in their paper cannot be justified because of lack of a comparison group. Regression to the mean is a common confound in test-retest studies such as presented in their paper. Inclusion of a comparison group could be used to rule it out. Other minor issues are raised involving scaling and consistency in the data.

Paul C. Bernhardt
Frostburg State University
Frostburg, Maryland, USA
pcbernhardt@frostburg.edu
+1-301-687-4410
Comment on Nalliah and Allareddy (2014): Unwarranted conclusions

In the article “Weakest Students Benefit Most From a Customized Educational Experience for Generation Y Students”, Nalliah and Allareddy (2014) present evidence of efficacy of a unique instructional method. They state that students who scored relatively lower on a pre session test showed improvement on a post session test due to the use of the instructional method. I will show that this claim is unwarranted by showing that the data is best explained by regression to the mean effects and the lack of a control group to rule it out as an explanation.

Findings and Claims

Nalliah and Allareddy (2014) measured 65 third-year dental students on 21-question dental knowledge test computing a 100 point (percentage) score for each. An instructional exercise was then conducted followed by a post session test. The pre session test, instructional exercise, and post session test all occurred in a single day. They report as the key finding that the 14 students in their designated low score group averaged higher scores on the post session test (Change in scores: $M = +9.5$, $Median = +14.3$). They also report that “surprisingly” (p. 8) the 31 students in their designated high score group yielded lower scores on the post session test (Change in scores: $M = –6.31$, $Median = –4.76$).

A causal conclusion is indicated in the title by the word “benefit” abstract by “benefitted” and in the primary conclusion by the word “improved”, quoting:

One of the objectives of this educational experience was to go through a straightforward case as a class. This would enable students to become familiar with the process of developing and defending their treatment plan with a Senior Tutor and help to improve their knowledge of basic cases which could improve their future learning experiences at HSDM. An unexpected benefit was that weaker students improved their pre-versus post-session test scores the most. (Nailliah & Allareaddy, 2014, p. 7)
Concern Regarding Regression to the Mean

The conclusion indicates the observed changes in scores from pre session test to post session test are due to the unique instructional method. However, the data is strongly suggestive of regression to the mean occurring and the design of the study cannot rule out that explanation. The drop in scores for the high scoring group and non-significant changes for the middling scoring students are an alert that regression to the mean is operating. If the study had a comparison group that did not experience the unique instructional method it would be possible to infer any differences between the pre and post session tests could be causally related to the unique instructional method. But, as the study is described in Nalliah and Allareddy (2014), regression to the mean is the more parsimonious explanation because of the ubiquity of the effect in test-retest situations.

A review of regression to the mean is probably appropriate to clarify the concern. These issues are covered in numerous textbooks. As an example, Rosenthal and Rosnow (2008, p. 210) describe regression to the mean as the mathematical consequence of any test-retest situation in which the correlation between the two tests is less than a perfect 1. In practice, all test-retest situations have $r < 1$ assuring regression to the mean occurs. The underlying reason for this is based in classical testing theory, also described in numerous textbooks. As an example, Morling (2015, pp. 312-315) describes that all measured scores are the sum of a true score value and an error in measuring that value. The differing error component for the first and second test is what reduces the correlation beyond any correlation between the unknown true scores. That error is also the reason underlying regression to the mean. The error in measurement is due to any number of randomly occurring factors ranging from moment-to-moment attentiveness of test-takers to luck in guessing on questions. Low scores are much more likely to have a negative error component from the accumulation of several random factors weighing on them compared to the
true score intended to be measured. Similarly, high scores are much more likely to have a positive error component from the accumulation of several random factors boosting them compared to the true score intended to be measured. On a second measurement, the several combining random factors are unlikely to all be repeated and accumulate in the same direction. Therefore, on the second measurement, the extreme low score will usually move higher while the extreme high score will usually move lower. That pattern is the one seen in Nalliah and Allareddy (2014). Figure 1 is a scatterplot of their data from the showing the change in score graphed against the pre session test score (Pearson $r = .522$, $p < .001$).

![Figure 1: Change in score by pre session test score and grouping](image)

Further evidence for regression to the mean operating in this study is that the standard deviation for the post session test is consistently larger than for the pre session test. For the middle two groups, it is inevitable because all students had the same score for the pre session test,
a standard deviation of zero. For the low scoring group the standard deviation for the pre session
test and post session test were, respectively, $SD = 7.64$ and $SD = 13.95$. For the high scoring
group the standard deviation for the pre session test and post session test were, respectively, $SD = 
2.26$ and $SD = 12.98$. That large spread in data from pre session test to post session test for
extreme groups is the expected consequence of regression to the mean.

Since regression to the mean must be assumed to operate in all test-retest studies a
comparison group must be included to have a chance to establish any causal effect of an
intervention (Morling, 2015, pp. 312-315). In the context of Nalliah and Allareddy (2014), a
comparison group’s purpose would be to quantify the regression to the mean effect so that any
greater effect could be judged to be due to the unique instructional method. But, this is the feature
critically lacking. Therefore, the causal claim in Nalliah and Allareddy cannot be supported and,
in my opinion, the paper should not have been published as a peer-reviewed paper in PeerJ with
language that implies causation.

Other Minor Concerns

**Score Rescaling.** I have minor concerns in the scaling of the test scores, some statistical
calculations, and in the supplemental raw data files. Nalliah and Allareddy (2014) rescaled their
21-point test to percentage scores. In this case, the scaling of the test score variables to
percentages tends to obscure more than it reveals, in my opinion. While a percentage score will
usually communicate to readers more clearly than the 21-point scale of the original test, in this
case it tends to communicate a greater degree of change in scores than is really happening. For
instance, the mean change score for low scoring students was 9.5%, which is only 2.0 questions
on the test. The mean change score for the high scoring students was –6.31%, which is –1.3
questions on the test. Changes that rely on the outcomes of so few questions can be reasonably
judged as fragile and cautious interpretation is appropriate. A further effect of this rescaling of
the test scores is that the nature of the groupings is difficult to determine. Ranges of scores such as $> 90$ to $\leq 95$ tend to connote scores scattered among that range. However, a careful reading of the tables and a glance at Fig. 1 above shows, all those points are only a single pre session test score for students: those who were correct on 19 of the 21 questions and scored a 90.48%. This would be more directly represented if the authors had stayed with the 21 point scale. These are not, however, critical failings and should not have resulted in rejection of the paper, though a cautionary note may have been warranted.

**Median Change.** The changes in scores for each group were also described in terms of changes the median (after describing them in terms of the mean). For example, “The median scores increased by 14.3 points within this group” (Nalliah & Allareddy, 2014, p. 5). The phrasing is unclear if it is the median of the change scores or if it is the change from the median of the pre session test scores to the median of the post session test scores. Examination of the raw data and tables in the paper reveals that it was the change in from the median of the pre session test scores to the median of the post session test scores. This makes it difficult to determine the real magnitude of change, and tends to make it appear larger than may be warranted. For example, in the low scoring group the median of the change scores was +9.52, lower than the change in medians of +14.3 indicated in the paper. For the high group, the median of the change scores was 0 (zero), a smaller degree of change than the –4.76 indicated in the paper.

**Data Error.** Another concern is notable in the data file offered in the supplemental information (DOI: 10.7717/peerj.682/supp-1). It has an apparent error that carried over into the calculations for the paper. The last case in the data shows the number of correct questions for the post session test as 19, but the percentage is “33”. Furthermore, 33 is shown as the minimum value for the post session test in the tables in the paper. It suggests one student score 20/21 on the pre session test and 7/21 on the post session test, a very unlikely outcome. The consequence of
this is to lower the means and medians, and increase standard deviations. Of course, it is not possible for me to definitively determine whether the error was in the recording of the test raw scores or in recording/calculating the percentage. However, the percentage value seems incorrect because the scores for each of four questions groups are in the data file and those scores total to 19. Therefore, I did all work for this comment as if the percentage for that particular item was 90.48. This has a minor effect on the findings as shown in Table 1, below, which attempts to replicate Tables 1 and 5 of Nalliah and Allareddy (2014). There is no change in the statistical findings due to the recalculation shown in Table 1. It should be noted that this issue does not arise if the rescaling of the data from number of correct questions to percentages is not done, as suggested above (assuming I have correctly determined the nature of the error).

**Table 1:**

Revised data based on presumably corrected data file for Nalliah and Allareddy (2014)

<table>
<thead>
<tr>
<th></th>
<th>All Students (N = 65)</th>
<th>Pre session test &gt; 95 Students (N = 31) †</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre session test score</td>
<td>Post session test score</td>
</tr>
<tr>
<td>Mean</td>
<td>90.26</td>
<td>*90.84</td>
</tr>
<tr>
<td>Std. Deviation</td>
<td>10.39</td>
<td>*9.68</td>
</tr>
<tr>
<td>Minimum</td>
<td>57.14</td>
<td>*52.38</td>
</tr>
<tr>
<td>Maximum</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Percentiles 25</td>
<td>85.71</td>
<td>*90.48</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>90.48</td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>90.48</td>
</tr>
</tbody>
</table>

* Values changed due to presumably corrected data from Table 1 of Nalliah & Allareddy (2014)

† Values changed due to presumably corrected data from Table 5 of Nalliah & Allareddy (2014)

**Conclusions**

In my opinion, the described paper should not have been accepted as a peer-reviewed publication due to the unsupported causal conclusion indicated in the paper, abstract, and title. It could have been accepted as a peer reviewed paper, in my judgment, if the conclusion was...
described as an *association* between the change scores and the relative level of pre session test performance, noting a unique instructional method was attempted and that regression to the mean could not be ruled out as possible influencing factor. The minor problems were not fatal flaws that should derail peer-reviewed publication, though they should probably have been caught in the peer review process. Lacking viewing of the peer reviewers’ comments, I cannot determine if any reviewer commented on these issues.
