Review of: "Interactive e-Contents: A Novel Gamification Approach for Students' Satisfaction"

José Antonio González Alastrué[1]

1 Universidad Politécnica de Cataluna

Potential competing interests: No potential competing interests to declare.

Some parts of the statistical analysis seem wrong to me, and I strongly ask the authors to review and correct them.

Section “FIndings”: – so nonparametric methods such as binomial test were used to analyze observations --. You don't need to use the Kolmogorov-Smirnov test to check if your variables are Normal or not. Your variables take 1-2-3-4-5 values, so it's crystal clear that they are not. Besides, the binomial test is a parametric method, it does not work with the Normal distribution but it does with the binomial distribution. Please rewrite the sentences.

Later: – values greater than 3 are considered as success and values smaller than and equal to 3 are interpreted as failure --. The authors consider hypothesis testing to check correctness of the hypotheses, although I have numerous doubts regarding the method. 1) Multiplicity of the tests: if testing is applied repeatedly, the type I risk is greatly increased; 2) Obscure parameter: what is “p”? Is it the probability to obtain a success, that is, an answer greater than 3 in some item? Clarify it to readers. Incidentally, Greek letters like **π** are tipically used for that purpose; 3) Mysterious 0.6 value: why this value? No reasons at all for this choice; 4) Wrong alternative hypothesis; regardless of the critical value (0.6) employed in the null hypothesis, the alternative hypothesis should not be “different”, but only “greater than”: would you accept that some item is correct since its success probabillity is 0, a value clearly “different” from 0.6? 5) Inconsistent presentation of results: is it a typo mistake, or the authors changed their mind? Table 5 (actually Table 6) shows participant distribution according to either “3 ≥” or “3 <”, although they stated that a success is a value strictly greater than 3; 6) Conclusions from the testings are totally wrong; assuming that we obtained for PU variable 4 “good” answers “3 ≥” and 31 “bad” answers “3 <”, the output for a formal binomial test is something like that:

> *> binom.test(4,35,0.6, alt="gre")*
>
> *Exact binomial test*
>
> *data: 4 and 35*
> *number of successes = 4, number of trials = 35, p-value = 1*
> *alternative hypothesis: true probability of success is greater than 0.6*
> *95 percent confidence interval:*
> *0.03999289 1.00000000*
> *sample estimates:*

> *probability of success*
>
> *0.1142857*

The P-value is not close to zero, but close to 1, meaning that in no way you would abandon your null hypothesis "$\pi \leq 0.6$" in favor of the alternative "$\pi > 0.6$", which is the goal of this investigation. Actually, it's still worse, since a naif estimation of the real probability of success for PU would give you the interval (0.032, 0.267), really far from the aimed 0.6. The mistake is carried on the rest of variables.

Of course, I think that you simply reversed the groups. Based on true Table 5, the mean average in every question is rather high (around 4), which is consistent with large proportions of "4" and "5" answers. So luckily it seems to me that the results obtained are rather positive. My advise: drop all these testings and give 95% confidence intervals for the success probabilities of the different items. Some of them will deceive your expectations (for instance, Challenge returns a CI 0.54 – 0.85, which includes the strange 0.6 value), but I'm convinced that the readers will appreciate it and, besides, allows you to avoid the use of mysterious threesholds like 0.6.

Additionally, it is a bit worrying that few reviewers point out big errors like those in the "Results" section. I have the feeling that everyone is trusting completely on the statistical analysis of the authors in any paper, and have no curiosity at all to check whether the analysis is correct, which is the foundations of the research. No research is valid without a thorough methodology and a correct data analysis, clearly. I acknowledge that I haven't checked exhaustively all results in this one, and maybe some other parts are failing, so I ask others to complete the task and positively criticize the validation of the instrument presented by the authors, which undoubtedly can be useful in the future if its usefulness is rigorously demonstrated.