On page 1 the authors cite the incredible figure of $200 trillion of additional revenue to the worlds GDP. I would love to see how this was arrived at. Is this for a single year? A decade?

Later on page 1 they give a too brief description of the test. They must include how the children were sampled. If they were not part of a rigorously random sample all of the subsequent results they cite are of no real value.

The results they report later showing that girls do better than boys in both math and reading in Qatar and Jordon are hard to understand given the education challenges faced by women in Islamic countries. Either our perception of such countries is far wrong, or the girls that were drawn into the sample were not representative. The authors should provide evidence for the interpretation they support. The nonresponse rates for each country would be a big help in understanding how credible are the conclusions drawn.

When they report scores both the mean and the standard deviation would be helpful for the reader to get some notion of scale.

Significance tests are rarely of any value for studies like this since the sample sizes are so big that any difference is statistically significant. What is important is what is called “clinical significance” in medicine – how big a difference must be for us to care.

Figures would be helped by headings emphasizing the meaning of the direction – so on top of Figure 1 on the top left an arrow pointing left saying “In favor of girls and another on the right pointing to the right saying “in favor of boys.”

I note in section 2 an obvious explanation – if they substituted “number of years of schooling” for age starting school it is plausible that the effect they highlight goes away.

Last, this is a two-factor design – countries by years at school (or year started) and ought to be summarized as factorial designs are summarized (two-way ANOVA with row and column effects, etc.) . Having a big multi-part plot is easy to understand.

\*\*\*\*\*\*\*

This is a very long read, and while it presents some interesting results there are a lot of quite concerning issues with regards to some fundamentals of statistical analysis. The main one is that in places it feels an awful lot like a 'fishing expedition' - that is, it reads as if someone has had access to a large dataset and kept playing with it until interesting results popped out. I've highlighted one particularly worrying example below with regards to "Albania is special". I think this can be dealt with fairly easily by either an opening paragraph that explains why they looked at the variables they did, or beginning each section with "Our next hypothesis was that...". In the section on parents, for example, the authors point out that there is considerable literature on the issue, and so it is perfectly reasonable for them to have investigated it. On the other hand, no reason is given for why they looked at Albania, or the number of televisions.

A second issue is that most of the results are at best findings of an association between two variables, but are in places presented as if there is evidence of a causal relationship. I have discussed one or two specifically below, but a good example is in the closing paragraphs where the authors observe "It is interesing [sic] that the more TVs in the household in developed countries leads to lower math scores on average." - this has not been shown at all. They have shown that in developed countries students from households with more TVs tend to have lower math scores on average, but there is no evidence that TVs are \*causing\* the lower math scores (as is implied by "leads to" in the preceding quote). If we are going to publish this it needs considerable work to acknowledge these issues, clarifying where appropriate that these are only associations and that we cannot make such claims about causality. The authors could perhaps propose reasons for why there might be a causal link, or why there might not be.

Thirdly, and again I occasionally address this specifically below, is that much of the results are seemingly presented as having looked at plots and spotted trends by eye. It may be that the authors did conduct some formal tests, but these are seldom mentioned.

Finally, there are obviously far too many figures in the document, and many of them are too complicated for publication. The authors need to be much more selective in which figures they think are worth including.

Here are rather a lot of specific points, which get somewhat more general as we go along.

"The size of the dot indicates the proportion of boys in the sample size." - while this is perhaps obvious, I don't think it would hurt to clarify this even more by "Larger sized dots indicate samples with a larger proportion of boys."

The paragraph "Although not shown here...more important than your demographic." - the point of this paragraph is to observe that there is between-individual variability that dominates between-gender variability, but I think focusing on the top math score in each country is a touch misguided. The top score in each country is largely irrelevant to the discussion, and I think the point of individual-level variability can be made without bringing up this specific example which potentially gives the impression of it being an important statistic, when it's really just an anecdotal example.

"For each country, we have computed the average math score by age that the children started school. In addition, for each country we computed the median age that children in the study started school." - the second of these should really come first, as the much simpler statistic. I would perhaps go with "For each country we computed the median age that children in the study started school, as well as the average math score by age that the child started school."

Were school start ages only available in whole years? That is certainly the impression of the graphs on page 7 (which, I presume, will have an explanatory caption?). I think it would be good if the authors could acknowledge this, as if you're looking at an age range of 4-8 a difference of a few months is quite considerable.

"School start age matters. Children need to start school in the 4-6 age range to expect better math scores, on average." This seems to be based on simply looking at the graphs, if there is a more rigorous analysis underpinning this claim then I think it should at least be mentioned (even if not in any great detail), otherwise this looks a bit unfounded, in which case this sentence can be dropped without too great a loss - it is fine to allow the reader to come to their own conclusion, I think.

The 'shortages' plots need better labelling - the meanings of 0, 1, 2 and 3 have to be inferred from the line plot below. Also these charts should all have the same y-axis scale (i.e. they should all go up to over 9000) if we are expected to compare them (as implied by "teacher shortage is less comman [sic] than material shortages"). I wonder if these could instead go on one stacked bar chart instead?

Is there a reason why the 'to some extent' and 'a lot' groups (I assume the 2 and 3 in the bar plots?) were combined in the line plot? This perhaps needs addressing in a future caption for that figure. That said, I think this plot is possibly too complicated for inclusion, there is a lot going on but it is difficult to infer anything useful from it. Instead any patterns would be best addressed in the text, and perhaps with one or two specific countries highlighted for illustration if deemed necessary.

"Scores decrease by 10-20 points onaverage [sic] in countries like Australia, China, Uruguay, Turkey, UAE, USA." - 10-20 points per what? Per unit of teacher/computer/library shortage? This needs clarifying.

"Teacher shortages appear to have impact in a few countries like Greece, Slovakia, and Serbia." - this sentence doesn't really make sense to me. What sort of impact? Can it be quantified?

"For each country we have provided loess-smoothed curves" - a bracket here explaing what a loess curve is in plain English would help avoid alienating readers, e.g. "For each country we have provided loess-smoothed curves (approximate 'lines of best fit')" perhaps.

Section 4 on 'time spent out of study'. This reads as if the conclusion is that it is spending time out of school studying that causes improved math scores. The authors need to acknowledge that obviously correlation does not imply causation, as they do not seem to have taken any statistical steps to look for anything beyond an associative relationship. The sentence "The pattern suggests decreasing benefits to increased out of school study time after 10 hours." is the main problem, as it suggests the study time is of benefit, when it could simply be that more able children are more likely to study outside of school.

The parents boxplot (section 5a) should not be included.

The plot in 5b is also far, far, too busy for the magazine.

Section 6: Albania is special. This is a mildly concerning section. Did the authors have any a priori reason to seek out Albania specifically? If not, then this could just be due to chance. They don't seem to have conducted any sort of formal statistical test here (although even if they had, the large number of possible combinations of countries and metrics means that it would be surprising if there \*wasn't\* a country that seemed unusual on at least some of them).

On televisions: "Because there were so (relatively) few observations of "zero", for this investigation, the "zero" and "one" categories were combined to form a "one or less" category." - this seems questionable to me. If TV ownership is taken as a proxy for e.g. wealth, then I think there's an important fundamental distinction between owning zero televisions and owning one television. Just because there aren't many with zero doesn't mean this is a distinction that can be overlooked.

The final figure is indeed a complex diagram, and I think to justify its inclusion it requires a much more careful explanation. The current paragraph that addresses it is effectively just a set of bullet points and is not very accessible.

Section 7: why was a linear model chosen? Was there a good reason to assume this would be a good fit? Did the authors check model diagnostics to verify the underlying assumptions weren't being violated? These are points that need addressing for the reader to be able to be confident that the analysis was at least reasonable.

Non-stats/typos:

"Five countries have gender gaps in favor of girls, and even more surprising, these are middle eastern or Muslim countries." - I think the "surprising" here really needs to be qualified, while obviously not the intention, at present this could be seen as offensive.

"Britain, and its past colonies, Australia, Canada and New Zealand" - while this is literally accurate, I'm fairly sure one should try and avoid using the term 'colonies' these days...

"Colombia has the biggest gender gap, about 30 points on average boys do better than girls." - this should read "on average, boys do about 30 points better than girls".

Data plural/singular.

"It should be noted that this is small, though, because this 30 points is out of a possible 1000 points." - that last 'points' is superfluous