**Reviewer #1**: Overall  
The authors describe challenges and vulnerabilities of causal inference methods such as MR and CIT that employ instrumental variables in the attempt to determine whether the true model is causal or reverse causal in the presence of measurement error. Conventional MR assumes the direction of causality is known, thus the authors propose an approach for relaxing this assumption and letting the data suggest a direction. This is achieved by comparing ro(gx) to ro(gy) using Steiger’s Z-test. The authors apply this approach to a dataset that includes SNPs, DNA methylation, and gene expression to determine whether variation in methylation affects gene expression or the reverse.

**Thank you to the reviewer for the helpful comments on this manuscript**

Major Issues  
On page 5, second paragraph of MR causal test, “The method that we will now describe is designed to distinguish between two models, x → y or y → x. Unlike the CIT framework, this approach cannot infer if the true model is x <- g -> y.” Also, in the Discussion (pg 11, second paragraph) the authors state that “in this work we assumed that pleiotropy was not present.” However, isn’t the non-causal model depicted in Figure 5b, just that, pleiotropy? The instrument g is affecting the outcome y through a mechanism other than the exposure x.

**The reviewer is correct, we have not been clear about what we have evaluated. The Steiger test does assume that there is no horizontal pleiotropy, but nevertheless we tested both the CIT and Steiger tests to evaluate their performances when horizontal pleiotropy was present, as shown in Figure 5b. We have clarified this throughout the manuscript.**  
  
By extending the MR approach, the authors have relaxed the assumptions, but it is not clear exactly what the new assumptions are and the consequences of violating those assumptions. If an assumption of the proposed MR approach is that pleiotropy is not present, then it might mislead some potential users to apply the approach to data explicitly simulated under conditions that violate the assumptions, i.e., the non-causal model described on pg 7 in section Simulations and depicted in Figure 5b. If the authors are suggesting that absence of pleiotropy is a weak assumption that can be ignored with minimal repercussions, then this should be stated in the text and pleiotropy explored explicitly in simulations.

**Thank you for this helpful comment. Actually we believe that absence of horizontal pleiotropy is quite a strong assumption and it leads to issues with both types of methods, and this is an important theme of the paper – that violating these (often untestable) assumptions problems are potentially quite severe and therefore caution is advised. We have tried to make much clearer the assumptions of the Steiger method.**

If absence of pleiotropy is an assumption of the proposed MR approach, under what conditions are investigators likely to know that x <- g -> y could not be true but x → y or y → x could be true? For example, in the real data analysis, how is it known that a single SNP could not affect both a CpG site and expression of a gene through independent mechanisms? Wouldn’t this case be in violation of the proposed MR approach assumptions?

**Yes we completely agree. The MR approach using a single instrument cannot reliably distinguish between the horizontal pleiotropy or a causal model, and this has been pointed out in a number of recent papers (Zhu et al 2016 (https://www.nature.com/ng/journal/v48/n5/abs/ng.3538.html), Hannon et al 2017 (http://www.cell.com/ajhg/abstract/S0002-9297(17)30158-1), Richardson et al 2017 (http://www.biorxiv.org/content/early/2017/04/29/132019), Chun et al 2017 (http://www.nature.com/ng/journal/v49/n4/full/ng.3795.html)). One way in which the Steiger test helps here is that for a large proportion of horizontal pleiotropy values the Steiger test will not return a ‘significant’ result because the cor(g,x) and cor(g,y) may not be sufficiently different. Nevertheless it is important to note that mediation methods also might perform quite poorly in separating causal from pleiotropic models (Figure 5), which ties into the theme of the paper – there is no method that should be considered particularly trustworthy when assumptions are typically untestable.**

Figure 3. Comparing methods with a power comparison is only meaningful if type I error is comparable. The problem here is how to define the null scenario since it includes many possible conditions that include dependencies. There are often trade-offs between type I error and power. The power comparison would be more meaningful to me if the issues of type I error were more fully addressed both conceptually and using simulations. The authors should make clear what conditions they are considering as legitimate under the null hypothesis of no mediation and compare methods under a range of those conditions. This will likely depend on the assumptions of the method since the authors would presumably caution against using the approach if the assumptions are not met.

**We thank the reviewer for this important point. We realise that the FDR for the CIT method is much lower than expected by chance, likely because a number of different null hypotheses need to be rejected jointly. As a consequence, Figure 3 is not a fair comparison.**

Pg 7, second paragraph of Simulations section. “Similarly in the non-causal model:”. Actually, the non-causal model is not as similar to the causal model as one might initially assume, due to the introduction of a hidden variable “u”. To simulate the model depicted in Figure 5b with measurement error, similar to the causal model, I would expect y = a + bx + e of the causal model to simply be replaced with y = a + bg + e. However, rather than just do that, the authors have introduced a hidden (unmeasured) variable, u, which affects both x and y. This model (without x0 and y0) was explored in Millstein et al. 2009 (see their Figure 4B). Millstein et al. pointed out that the model is mathematically equivalent to a model where a hidden variable, not affected by g, affects both x and y. In view of these issues, it would help readers evaluate the proposed approach if the authors, 1) include an assessment of the performances of the methods under the non-causal model without the hidden variable u, 2) make clear in the text and Figure 5b that there is a hidden variable, and 3) make clear that the hidden variable model is mathematically equivalent to another model where the hidden variable, u, is not affected by g.

**Thank you for bringing this to our attention. The non-causal model, as specified, is one that essentially represents horizontal pleiotropy, where the SNP influences both the exposure and the outcome but through independent pathways. The way it is formulated is actually functionally identical to the model that the reviewer proposes, the only difference being that the variance explained by g on x or y will be cor(g,u)2cor(u,x)2 and cor(g,u)2cor(u,y)2, respectively, rather than cor(g,x)2 or cor(g,y)2 directly. Here bux and buy are identical.**

**Following this prompt, we have now performed a more extensive exploration of the influence of hidden confounders on both the mediation and MR based approaches. This is now detailed in Appendix 3, with discussion throughout.**  
  
In the CIT paper (Millstein et al. 2009), they interpreted the outcome where Pvalue(cit, x -> y) < alpha and Pvalue(cit, y -> x) < alpha as ‘no call’. However, in this paper it is interpreted as (x <- g -> y). This difference in interpretation should be explained and justified in the text. Data simulated under the simple model representing (x <- g -> y), that is, g -> x and g -> y, will typically result in a non-significant CIT p-value (CIT p-value > alpha). It might make sense to classify outcomes of CIT into 1) causal, 2) reverse causal, and 3) not significant, and 4) no call.

Figure 2 should be modified to include both the proposed MR approach and CIT.  
  
Minor Issues  
  
Bottom of pg 6. “We find the ratio of the volume that agrees with the inferred direction of causality over the volume that disagrees with the inferred direction of causality.” This sentence could be more clearly stated. Do the authors mean, “We find the ratio of the volume that agrees with the inferred direction of causality [given no measurement error] over the volume that disagrees”?  
  
Figure 3, n = 10,000. Power is 1 for both scenarios across the entire range, so these plots are uninformative and can be excluded.  
  
Figure 3, n=1000. Upper plot. The CIT results appear to conflict with the results presented in Figure 2. In Figure 2 there is monotonically increasing power with increasing cor(x,x0) over the range of (0,0.4) but in Figure 3 power for CIT decreases over this range.  
  
Figure 3, n=1000. Upper plot. There appears to be a problem with the results presented in Figure 3. If cor(x, x0) = 0, then there is no information in the measured variable x0 about x or y presumably. Thus for an unbiased test, the true positive rate would be the type I error rate. The Figure 2 plot is consistent with this idea with –log10p very small with small correlation. However, in Figure 3, there is substantial power for both methods for the first pair of bars where cor(x, x0) = 0.  
  
Figure 4 caption (a). Visually, it appears as though more of the volume between the red plane and blue surface is below the surface, not above. Yet R = 4.4, which would seem to imply that substantially more volume between the blue and red surfaces is in the positive y-axis domain. Maybe the 3d rendering could be improved to more clearly convey the increased volume in the positive domain?  
  
Figure 5. The top x-axis should be clearly labeled and described.  
  
Figures 3 and 5a, n=1000 scenario. In figure 3, MR appears consistently more powerful for n=1000, but in Figure 5a for n=1000, CIT appears constantly more powerful. The apparent discrepancy should be clarified in the text. Also, if the CIT is more powerful over some region of the parameter space, while MR is more powerful over another region, this should be mentioned in the Discussion.  
  
**Reviewer #2**: This was a mostly interesting and informative manuscript on a topical subject. It would benefit from some clarifications in places.  
  
Specific comments:  
  
1. The abstract, author summary and a number of other places (especially in the earlier part of the manuscript) suggest that you will be considering a number of different mediation-based approaches and will be demonstrating that all of them fail (or perform worse than your proposed new method, MR Steiger) when there is measurement error. However, in the end you only really examine one method, the CIT. This fact needs to be made much clearer all along. I don't dispute your statements in the Discussion that one might expect other mediation-based approaches to have similar issues as CIT, but you do not actually demonstrate this (or show the extent of the problem for methods other than the CIT).  
  
2. I was a bit surprised you did not describe in detail early in the manuscript what seems to me the most commonly used MR method for trying to infer the direction of causality, namely to perform two MR analyses (with two different instruments, one for x and one for y) and use the inference from the 2 tests in some way. (Is this known as bi-directional MR?) You sort of allude to this later on when you talk about multiple instruments, but I would think it worth mentioning much earlier (and even including it in your comparison).  
  
3. Page 6: Your use of the ratio of the volume that agrees with the inferred direction over the volume that disagrees, as a measure of "reliability" would seem to assume that all values of measurement error (within the bounds) are a priori equally likely. Is that correct? Is it realistic?  
  
4. Page 7: You describe \alpha and \beta as the parameters that govern measurement error, but surely \epsilon is also another such parameter? Actually I think you mean to include it (i.e. you mean to say \epsilon, \alpha and \beta are parameters that represent measurement error) but the sentence becomes confused so that the reader erroneously sees two separate phrases:  
(a) \epsilon is normally distributed  
(b) \alpha and \beta are parameters that represent measurement error  
This should be clarified.  
  
5. Page 7: Actually you only seem to use \epsilon in your simulations - you assume no fixed displacement (\alpha=0) and no multiplying factor (\beta=1). Is this realistic? Would your results be different if you relaxed this assumption?  
  
6. Figure 3 shows true positive rates (power) - what about false positive rates (type 1 error)?  
  
7. Page 8: By "no measurement error in the outcome or exposure variables (\rho\_{x,x\_0=1)", do you need to also add "and \rho\_{y,y\_0=1)" ?  
  
8. Page 9: "the MR analysis is indeed liable" - do you mean "the MR Steiger analysis is indeed liable"? There are a few other places where you talk about MR analysis when I think you mean MR Steiger analysis (i.e. your new method) - it would make things clearer if you could be more precise.  
  
9. Page 10: Is CIT really "widely used in predominantly 'omit data"?  
  
10. Figure 1: Are either the legends, or the pictures, for part (a) and (b) swapped? It seems to me that (a) represents the situation where gene expression causes methylation, and (b) represents the situation where methylation causes gene expression, but the figure legend has them described as the other way round.  
  
**Reviewer #3**: My main concern with the presented results is that the authors discuss imprecisely measured traits, but they do not reflect on imprecisely defined traits. Do their methods address these related issues too, and if not what are the implications for researchers considering their methods or prior methods?  
  
Additional comments:  
1. Abstract: "...and that increasing sample sizes has the adverse effect of increasing confidence in the wrong answer." This seems like such an obvious sentiment to me for a methods paper; it may be helpful for the broader audience but if this were to appear in a statistics/methods journal I would recommend excluding it.  
2. The first paragraph of the introduction overstates IV methods' superiority to other techniques. IV methods overcome baseline confounding only (not all confounding or all definitions of reverse causation) and come in exchange for further assumptions.  
3. Page 4, the authors note they assume no measurement error in the SNP (more specifically, they also seem to assume the measured SNP is causal). Can the authors expand upon how much this limits the usability or interpretation of their methods?  
4. The available methods and the authors' developments all seem to heavily rely on a constant treatment effect assumption. Is this true? If yes, can the authors comment on the usability or interpretation of applications of these methods in realistic settings? (Note of course that a constant treatment effect is biologically implausible in many settings.)  
5. I found the final paragraph of the main text a refreshing and sobering conclusion. I wonder if some of these sentiments should be better foreshadowed in the introduction.  
6. For methodology readers, it may be better to turn some of the heuristic drawings in the figures into formal mathematical graphs (e.g., causal directed acyclic graphs).