# Visualising two approaches to explore reliability-power relationships

Sam Parsons[1]

[1] University of Oxford

## Author Note

I would like to thank Anne-Wil Kruijt for her input on the original code used to generate Figure 1 in an earlier version of this paper. I would also like to thank Eda Tipura and Annabel Songo for their useful feedback on an earlier version of this paper.

The code used to run the simulations and generate this paper can be found in the OSF repository for this paper https://osf.io/msk2w/

The author is currently supported by a grant from the Economic and Social Research Council [grant ref: ES/R004285/1]

Correspondence concerning this article should be addressed to Sam Parsons, Department of Experimental Psychology, University of Oxford, New Radcliffe House, Radcliffe Observatory Quarter, Oxford, OX2 6AE. E-mail: sam.parsons@psy.ox.ac.uk

Abstract

The relationship between measurement reliability and statistical power is a complex one. Where reliability is defined by classical test theory as the proportion of 'true' variance to total variance (the sum of true score and error variance), power is only functionally related to total variance. Therefore, to explore direct relationships between reliability and power, one must hold either true-score variance or error variance constant while varying the other. Here, visualisations are used to illustrate the reliability-power relationship under conditions of fixed true-score variance and fixed error variance. From these visualisations, conceptual distinctions between fixing true-score or error variance can be raised. Namely, when true-score variance is fixed, low reliability (and low power) suggests a true effect may be hidden by error. Whereas, when error variance is fixed, high reliability (and low power) may simply suggest a very small effect. I raise several observations I hope will be useful in considering the utility of measurement reliability and it's relationship to effect sizes and statistical power.

*Keywords:* reliability, statistical power

Word count: 2988

# Visualising two approaches to explore reliability-power relationships

As defined by classical test theory, observed-score variance, or "total" variance is the sum of "true-score" variance and "error" score variance (X = T + E). Measurement reliability is defined as the proportion of variance attributed to true-score relative to total variance (R = T / T+E). Statistical power is the probability of detecting a specified effect size, or greater, at a given sample size and alpha. Statistical power is functionally related to total variance, such that increased total variance results in a reduction in power. Thus, power is related to total variance, regardless of how it is split into true-score and error variance. It follows that reliability does not therefore have a direct functional relationship with power. That is, unless either true-score variance or error variance is held constant while the other is left to vary; as Zimmerman and Zumbo (2015) demonstrate clearly. By varying only true-score or only error variance; total variance varies in proportion and we can examine the resulting relationship between reliability and power. Briefly, when true variance is fixed, increases in error variance result in decreases in reliability and decreases in statistical power. In contrast, fixing error variance and increasing true variance results in increases in reliability, yet, this is accompanied by a decrease in power. Zimmerman and Zumbo (2015) provide an excellent coverage of the mathematical definitions underlying these relationships; and I do not aim to simply repeat them in this paper. Instead, I present these relationships visually comparing a hypothetical measure with zero, the research question being, simply, does the group differ significantly from zero on this measure? First, total variance is fixed while true-score and error variance alternate. Second, true-score variance is fixed, while error variance is increased. Third, error variance is fixed, while true-score variance is increased. For each approach I visualise simple simulations that illustrate the relationship between reliability and power. I aim to highlight some of the implications and conceptual differences between the approaches of fixing true-score or fixing error variance.

**Holding total observed variance constant**

To visualise the relationships between reliability and power Figure 1 presents a somewhat abnormal kind of plot. The proportion of the line in blue represents the proportion of "true-score" variance, while the proportion in red line represent the proportion of "error" variance. Total variance is therefore the total width of the line; the wider the line overall, the smaller the effect size. The reliability of the measure (true variance divided by total variance) is presented on the y axis. For the purposed of this visualisation, we might consider lines that do not cross zero on the x axis to be significantly different from zero. The left panel assumes constant total variance; the middle panel assumes constant true-score variance; and the right panel assumes constant error variance.

First, an illustration of no relationship between reliability and power. In the left panel of Figure 1, total variance is held constant. As the proportion of true score to error score variance reduces, the reliability of the measure also reduces. However, the effect size remains constant demonstrating that effect size is influenced by the total variance, not the proportion of that variance that is attributable to true-score and error variance.

[Insert Figure 1 here]

This is also demonstrated in Figure 2, in which power does not vary in relation to reliability. Figure 2 was generated by simulating datasets in which the error variance and true variance were alternated between 0 and 1, in .1 increments, while holding total variance constant at 1. The group mean was set to .3. Sample sizes of 20, 30, 40, 50, and 60 were generated. 20000 iterations of this 11*5 design were performed. Statistical power was then calculated as the percentage of significant results at an alpha of .05.

[Insert Figure 2 here]

**Increasing error variance**

An often employed method to examine the relationship between reliability and power is to manipulate reliability by increasing the amount of measurement error. For instance, LeBel and Paunonen (2011) used this procedure and concluded that reductions in reliability reduce the confidence we should place in our analyses and in statistical power (and later extend this to replicability of results). Similarly, Kanyongo and colleagues (Kanyongo, Brooks, Kyei-Blankison, & Gocmen, 2007) used a Monte Carlo simulation approach (MC2G software; http://www.ohiouniversityfaculty.com/brooksg/) to demonstrate that increasing measurement error (and therefore, reducing reliability) decreases power for a range of analyses, including; correlations, paired t-tests, independent t-tests, one-way ANOVA (three groups), Wilcoxon signed-rank tests, and Mann-Whitney-Wilcoxon tests. When applied to correlational analyses, the decline in statistical power (and increase in required sample sizes to achieve 80% power) even at "acceptable" levels of reliability is daunting (Parsons, 2018). It is particularly striking given that reliability and validity information often goes unreported (Flake, Pek, & Hehman, 2017; Gawronski, Deutsch, & Banse, 2011; Hussey & Hughes, 2018; Parsons, Kruijt, & Fox, 2019).

Figure 1 (middle panel) displays this visually by setting true variance at 1 and increasing error variance such that reliability decreases (from top to bottom) in steps of 0.1. For the sake of interpretation, we might interpret lines that cross zero as being non-significantly different from zero. With enough added error variance, even real effects will be hidden. We can see that with decreases in reliability comes a decrease in the observed effect size, as the total variance increases. This is perhaps the common understanding of reliability, that low reliability equates to greater error.

To illustrate the relationship between reliability and power under the assumption of fixed true-score variance, simulated data were generated following Lebel and Paunonen's

(2011) approach. Figure 3 presents this visually. True variance was set at 1, and error variance increased such that reliability decreased from 1 to .1 in steps of .1. The group average was set to .3. Sample sizes of 20, 30, 40, 50, and 60 per group were created. 20000 iterations of this 10*5 design were performed. Statistical power was calculated as the percentage of significant results at an alpha of .05. Note that this plot closely matches the one created by Lebel and Paunonen (2011, p. 576).

[Insert Figure 3 here]

**Increasing true-score variance**

The alternative approach to examine the relationship between reliability and power is to fix error variance and vary true variance (De Schryver, Hughes, Rosseel, & De Houwer, 2016; Zimmerman & Zumbo, 2015). In this case the reverse pattern to fixing true-score variance is observed; power decreases with increased reliability when error variance is held constant. Figure 1 (right panel) displays this visually by setting error variance at 1 and increasing true-score variance such that reliability increases (from top to bottom) in steps of 0.1. Here, with increased true-score variance comes increases in reliability co-occurring with decreases in the observed effect size. Thus, the more reliable measure here also coincides with the lowest power.

As with the previous approach, data were simulated to illustrate the relationship between reliability and power, assuming that error variance is fixed. Figure 4 presents this visually. In this case, error variance was set to 1 and true variance was increased such that reliability varied from 0 to .9 in steps of .1. The mean group score was set to .3. Sample sizes of 20, 30, 40, 50, and 60 were created. 20000 iterations of this 10*5 design were performed. Statistical power was calculated as the percentage of significant results at an alpha of .05. Note that this plot closely matches the one created by De Schryver et al. (2016,

p. 4). Here, we observe that power and reliability have an inverse relationship.

[Insert Figure 4 here]

**Interpreting different reliability-power relationships**

The relationship between measurement reliability and statistical power is complex, in so far as it is not a direct relationship. To examine a direct, functional, relationship between reliability and power, either true-score variance or error variance must be fixed, while the other is held constant. Following mathematical definitions is useful in examining the reliability-power relationship (Zimmerman & Zumbo, 2015). As Visualisations can be a useful ullustrative tool to aid understanding I aimed to supplement previous articles with visualisations of the reliability-power relationship with the aim of elucidating the implications of either assummption.

In part, generating these visualisations was prompted by discussions about the relevance of measure reliability in different research contexts. As noted by Hedge and colleagues (Hedge, Powell, & Sumner, 2017, p. 3) "the meanings of a 'reliable' task for experimental and correlational research are not only different, but can be opposite in this critical sense" due to the relative advantage or disadvantage of larger variation in the within-subject effect. Fixing true variance has led to recommendations that we must improve the psychometric properties of our measurements, so as not to render our results uninterpretable and to increase the replicability of our research (LeBel & Paunonen, 2011; also see, Hedge et al., 2017; Parsons, 2018; Parsons et al., 2019; Rodebaugh et al., 2016). In contrast, the approach of fixing error variance and increasing true-score variance has been related to the idea that experimentalists seek to maximise an experimental effect (and have greater statistical power) by recruiting more homogeneous samples, fitting with the aim of reducing true-score variance, which results in reductions in reliability (De Schryver et al.,

2016). How can we reconcile these opposing stances on measure reliability? I suggest that part of the reconcilation comes from discussing the conceptual distinctions resulting from holding either error or true-score variance constant.

## Conceptual distinctions between approaches to examine reliability-power relationships

The purpose of this paper is to highlight conceptually distinct rationales for the fact that low power can result from increasing true-score variance or from error variance (while the other is held constant), and thus from both very reliable and very unreliable measures. To conclude, I leave the reader with two observations based on the visualisations above that I have not yet seen articulated in discussions surrounding the relationship between reliability and power. I hope that these conceptual observations are useful in considering the utility of measurement reliability in experimental research.

First, as highlighted in figure 1; for any specified level of total variance, reliability can take any value. In this case, regardless of reliability; power remains constant despite reductions in the "meaning" of the scores on an individual differences level. While this reduction in reliability does not influence the power of the between group difference, it does influence correlational analyses with this measure as individuals become harder to rank order. Likewise, it would be difficult to demonstrate that an individual's score has changed above chance or to investigate questions of individual differences. Further, it has been recently argued that the upper bound of validity of a measure is it's reliability (Zuo, Xu, & Milham, 2019). From this perspective, it is much more difficult to assume that one is measuring the process of interest, even with a highly powered test showing a large effect (e.g. in Figure 1, right panel). Thus, while an effect can certainly be found, the reliability and validity of that measurement may still be suspect.

In these cases, although a highly powered systematic effect has been observed, we cannot have much confidence that this difference is in the process or construct of interest. If reliability is so low that the score is rendered near-meaningless, does it matter how much more power the test has or how much larger the effect size is? Is the measure even capturing the very thing that researchers were interested in in the first place? Highly reliable measurements have the benefit that the measure is less contaminated by error and more likely to capture the process, or construct, of interest. The only potential benefit of low reliability is of maximising effect sizes under the condition that a certain amount of measurement error is expected. However, this comes with the drawback of a less interpretable measure, and follows the assumption that if there were sufficient true-score variance for a reliable measure then there would be little-to-no actual effect to find in the first instance. In all cases, reducing measurement error is beneficial to the interpretability of scores as well as increasing power (assuming that this reduction in measurement error does not co-occur with an increase in true score variance in excess of this reduction). While experimentalists seek to maximise the effect of interest by recruiting more homogeneous groups, it does not follow that low reliability itself is beneficial, only that reducing total variance (e.g. via reducing measurement error) will maximise the effect observed.

Second, consider the widest lines in the middle and right panels of Figure 1. These contain the maximum total variance - and thus, the lowest power of all the tests. Each suggests an alternative interpretation for the relationship between reliability and (low) power. When error variance is increased (Figure 1, middle panel), the true effect is hidden due to measurement error. Whereas, when true variance is increased (Figure 1, right panel), the actual true effect is smaller. There is simply less of an effect to find. Therefore, we should *expect* low power in this case. Although total variance is the same, the composition into true-score and error variance leads to different reasons explaining the resulting low power. We should not think of the reliability-power relationship under either approach as being a direct result of reliability.

These considerations are not in disagreement with mathematically defined functional relationships between reliability and power. I do, however, believe that there are conceptual considerations that must be made when interpreting the relationship between reliability and power under either assumption. These conceptualisations have implications for some claims that have been made about the importance of reliability in experimental research. (De Schryver, 2018) provides an illustrating example. Consider two measures, A and B, with error variance of .20. Score A contains true-score variance of .80 (therefore, a reliability of .80), while score B contains true-score variance of .60 (therefore, a reliability of .75). Score A is therefore preferable based on solely on reliability. Yet, calculating Cohen's d - assuming a mean difference of .50 - reveals that the effect size for A is .50 and for B is .56 due to the greater total variance in score A. Therefore, to maximise the effect size observed, a researcher might seek to use measure B. This example again highlights the fact that reliability does not have a 1:1 relationship with statistical power. Selecting a measure based on the size of the effect it produces with the aim of increasing power may be a valid approach. However, my concern is that this approach might be interpreted as an indication that low reliability is either; not a problem, or may be beneficial under certain conditions. This simplistic conclusion misses the fact that it is total variance that is functionally related to power and not reliability per se. I also worry that demonstrations of this negative relationship has led some to conclude that low reliability is therefore "acceptable" for experimental designs, and by implication that we should not be concerned about the reliability of our measures. This thinking was part of what prompted the original creation of the visualisations in this paper.

The aim of this brief paper is to expand current examinations of the relationship between reliability and statistical power (e.g. Zimmerman & Zumbo, 2015). Visualising these relationships under separate assumptions of fixed true-score variance and fixed error variance suggests each method represents a conceptually distinct process. We might consider each approach separately in terms of obscuring a true effect (fixing true-score variance) or reducing the size of the effect (fixing error variance). The fact that; under the first

assumption power increases with improved reliability, whereas under the second assumption the opposite is observed, is less paradoxical when considering these alternative explanations. This in itself might help to avoid potential simplifications that low reliability is acceptable under certain circumstances and promote a more in depth consideration of the influence of measurement reliability on statistical power in a given study.

**References**

De Schryver, M. (2018). *A psychometric analysis of choice reaction time measures* (PhD thesis). Ghent University.

De Schryver, M., Hughes, S., Rosseel, Y., & De Houwer, J. (2016). Unreliable yet still replicable: A comment on lebel and paunonen (2011). *Frontiers in Psychology*, *6*(JAN), 1–8. https://doi.org/10.3389/fpsyg.2015.02039

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Gawronski, B., Deutsch, R., & Banse, R. (2011). Response interference tasks as indirect measures of automatic associations. In K. Klauer, C. Stahl, & A. Voss (Eds.), *Cognitive methods in social psychology* (pp. 78–123). New York, NY: Guilford.

Hedge, C., Powell, G., & Sumner, P. (2017). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*. https://doi.org/10.3758/s13428-017-0935-1

Hussey, I., & Hughes, S. (2018). Hidden invalidity among fifteen commonly used measures in social and personality psychology. https://doi.org/10.31234/osf.io/7rbfp

Kanyongo, G. Y., Brooks, G. P., Kyei-Blankison, L., & Gocmen, G. (2007). Reliability and statistical power: How measurement fallibility affects power and required sample sizes for several parametric and nonparametric statistics. *Journal of Modern Applied Statistical Methods*, *6*(1), 81–90. https://doi.org/10.22237/jmasm/1177992480

LeBel, E. P., & Paunonen, S. V. (2011). Sexy But Often Unreliable: The Impact of Unreliability on the Replicability of Experimental Findings With Implicit Measures. *Personality and Social Psychology Bulletin*, *37*(4), 570–583.

https://doi.org/10.1177/0146167211400619

Parsons, S. (2018). Ignoring measurement reliability is a real-life horror story [Blog post]. Retrieved from https://medium.com/@Sam_D_Parsons/ignoring-measurement-reliability-is-a-real-life-horror-story-b98a2517db26

Parsons, S., Kruijt, A.-w., & Fox, E. (2019). Psychological Science needs a standard practice of reporting the reliability of cognitive behavioural measurements, 1–25. https://doi.org/10.17605/OSF.IO/6KA9Z

Rodebaugh, T., Scullin, R., Langer, J., Dixon, D., Huppert, J., Bernstein, A., ... Lenze, E. (2016). Unreliability as a Threat to Understanding Psychopathology: The Cautionary Tale of Attentional Bias. *Journal of Abnormal Psychology*, *125*(6), 840–851. https://doi.org/10.1037/abn0000184

Zimmerman, D., & Zumbo, B. (2015). Resolving the Issue of How Reliability is Related to Statistical Power: Adhering to Mathematical Definitions. *Journal of Modern Applied Statistical Methods*, *14*(2), 9–26. https://doi.org/10.22237/jmasm/1446350640

Zuo, X.-N., Xu, T., & Milham, M. P. (2019). Harnessing reliability for neuroscience research. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-019-0655-x

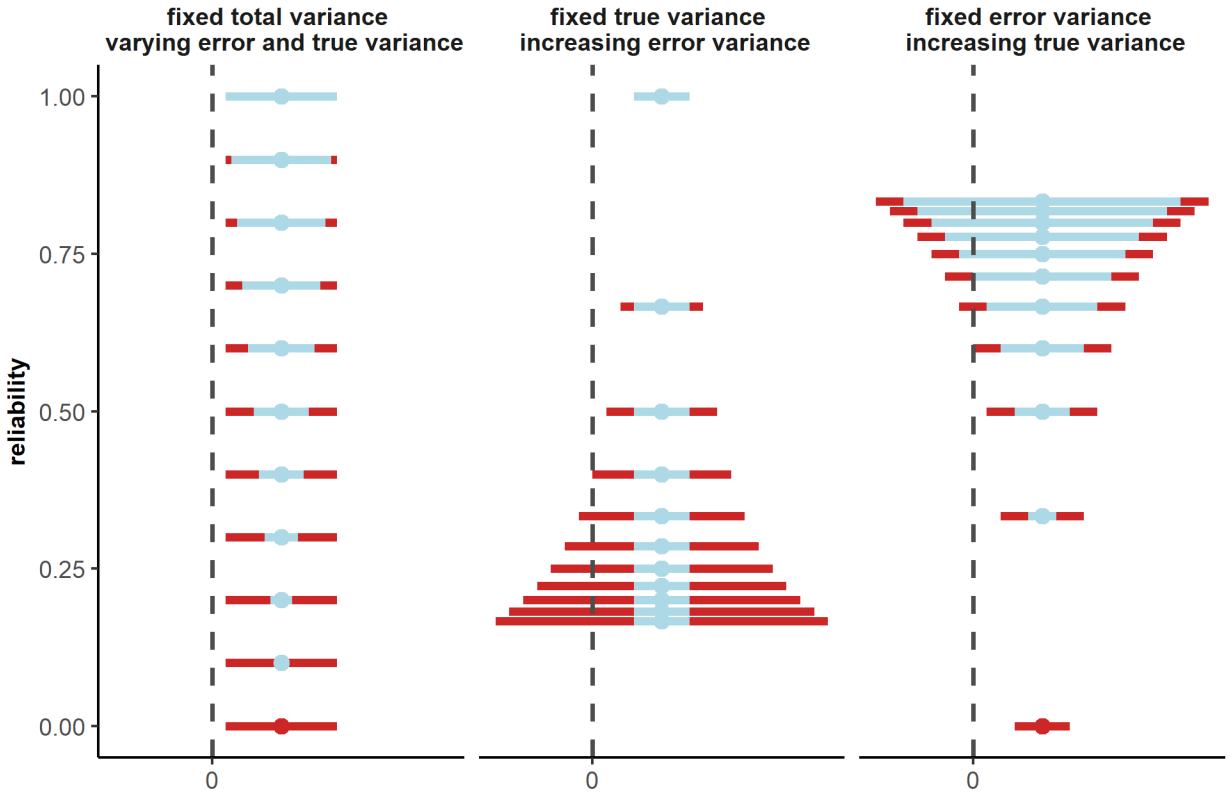


*Figure 1*. Visualising total variance decomposition into true-score variance (blue) and error variance (red). Reliability. The reliability (true variance / total variance) is indicated on the y axis. Left panel: total variance is held constant. Middle panel: true-score variance is held constant. Right panel: error variance is held constant

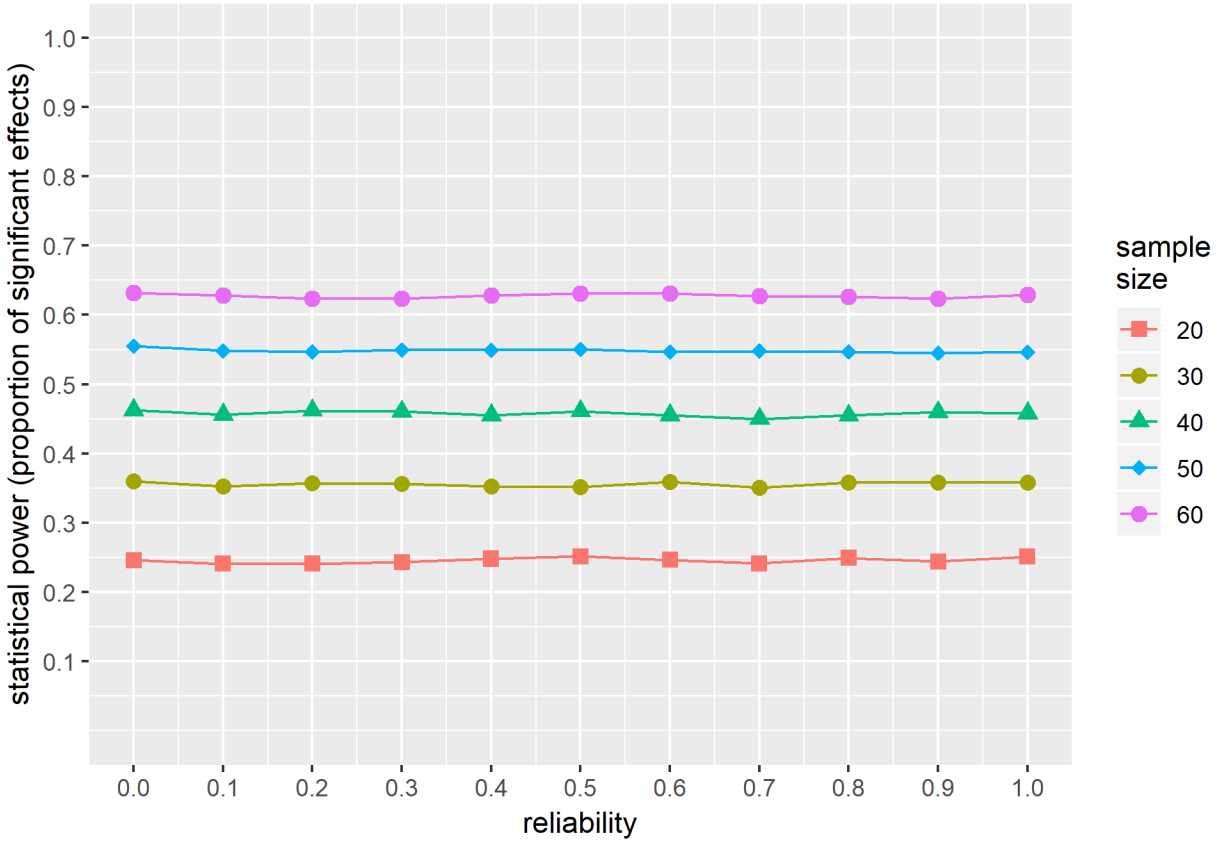


*Figure 2*. Observed probability of finding a significant result as a function of sample size and reliability where total variance is fixed (true-score and error variance alternate)

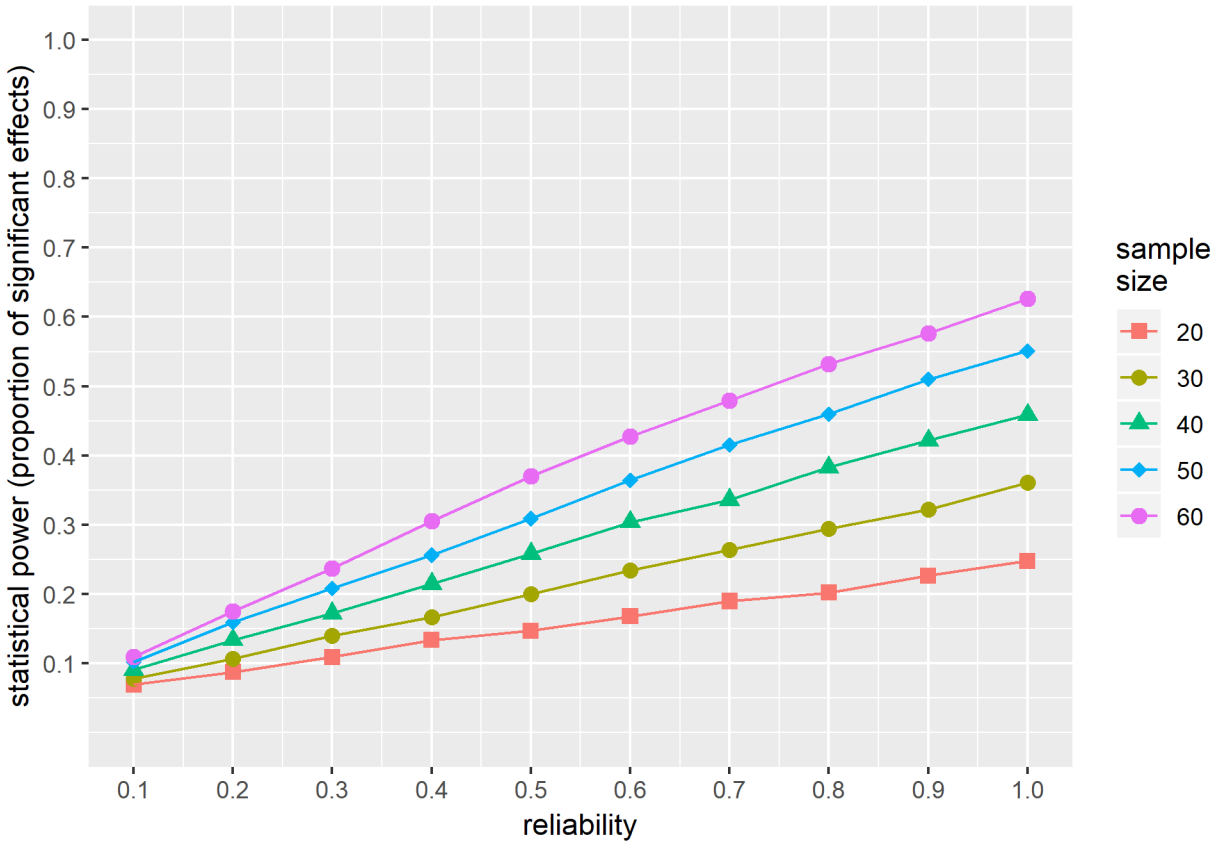


*Figure 3*. Observed probability of finding a significant result as a function of sample size and reliability where truescore is fixed and error variance is increased

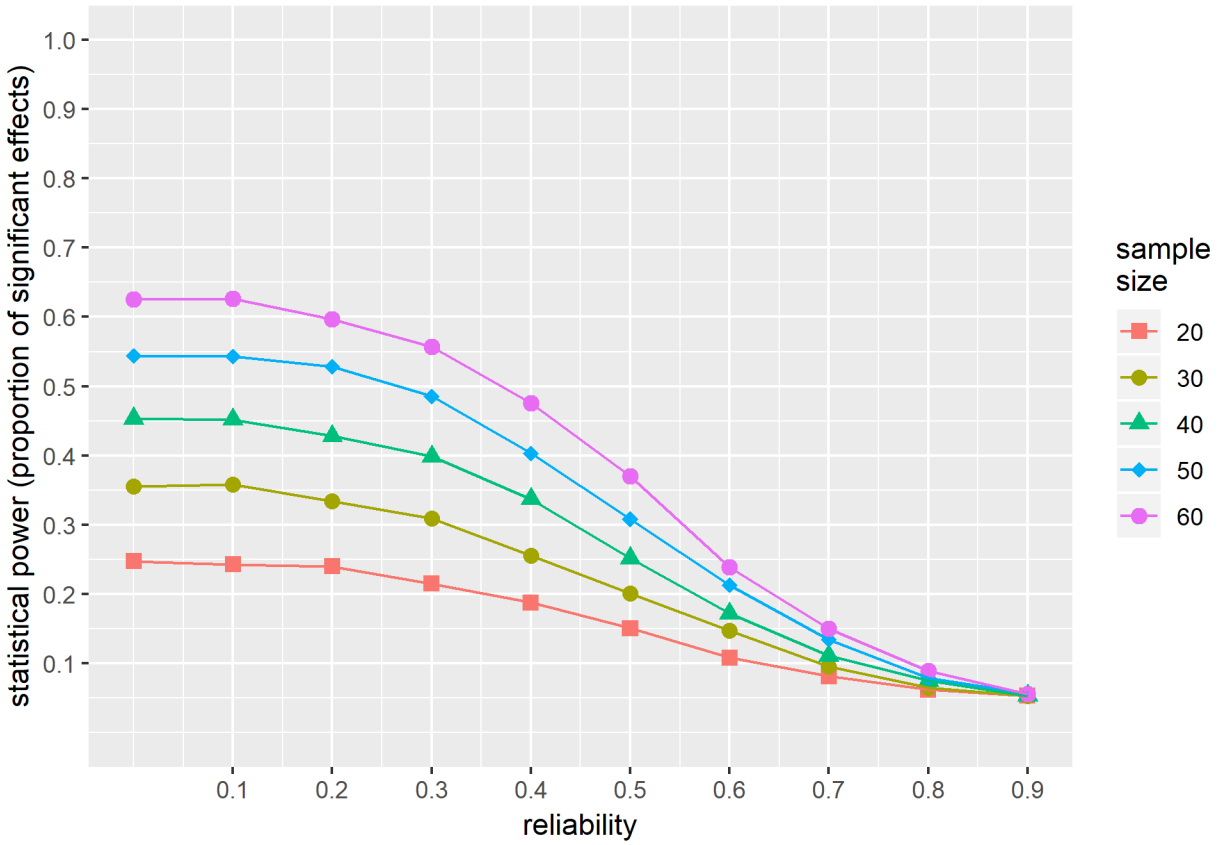


*Figure 4*. Observed probability of finding a significant result as a function of sample size and reliability where error is fixed and truescore variance is varied.