News, Comment and Reviews

A Critical Reply to Gorard on his Trustworthiness Matrix & Its Justification

Larry Brownstein

Gorard has produced what I think is a valuable instrument for assessing what he calls the trustworthiness of a research project or result. His discussion is separated into three parts, the justification for introducing this novel method of assessing trustworthiness, the trustworthiness matrix itself, and its interpretation. The trustworthiness matrix is itself in my view a worthwhile contribution to the field for which it is intended. My concerns are with what I conceive to be significant errors in his discussion. One lies with the justification for the creation of the matrix. The other is the interpretation of the matrix itself. With one alteration, the matrix itself I think is fine and useful and I hope will be widely disseminated and utilized.

I. The justification of the matrix

The central justification for Gorard creating his trustworthiness matrix is a critique of confidence intervals, which many might take for providing an insight into the trustworthiness of a piece of research and which Gorard is rightly concerned to refute. Gorard makes clear that his trustworthiness matrix is not to be considered as a substitute for an analysis employing confidence intervals. That is, confidence intervals should not be used to judge the trustworthiness of research results. Given the character of the quite different procedures associated with both of these ways of analyzing data, I would agree with Gorard's assessment.

While it is true, as Gorard points out, that confidence intervals have been misused and misinterpreted, some of the things he says about them are either false or do not make sense. But in addition, he uses Mathews (1998) to argue indirectly that confidence intervals can be dangerous in the sense that such interval analyses neither provide confidence in the results nor indicate a likelihood that the 'true' result will lie within the interval. I don't think many would dispute this. The misuse of any statistical procedure can be dangerous in the sense of providing confidence in findings that do not justify them, particularly in policy related fields. Gorard neglects to mention in the text, where it would be most apparent and useful to the reader, that Mathews' approach is Bayesian, specifically, a subjective Bayesian one. This is odd because he refers to a paper by Jaynes on Bayesian intervals in a seemingly unpublished paper as well as Mathews, so it is clear that Gorard is aware of the differences between the two kinds of interval analyses. Since it appears to be classical CIs that Gorard is concerned about, why reference Mathews?

It might assist the reader to know that confidence intervals are a classical statistical procedure invented by Jerzy Neyman (along with other concepts useful also to Bayesians such as Type I and Type II errors). Bayesian and classical statistics form two distinct and somewhat conflicting schools of statistical thinking. It should then be no surprise to discover Matthews being hostile to the deployment of classical confidence intervals. This is not perhaps the best source for Gorard to cite in support of his thesis for avoidance of these kinds of CIs.

You don't, however, get any sense that there might be potential conflict between Matthews' position and that of someone using classical confidence intervals from Gorard's few remarks. Moreover, in reading Matthew's paper, however, one finds that it is not really about confidence intervals at all but rather about the flaws of Fisherian hypothesis testing from a Bayesian perspective plus the nature of subjectivity in the assessment of research results. Jaynes (1976) makes clear what the issues are that lie behind these two interval assessments in his comparison of (classical) confidence intervals with Bayesian intervals (though he does it via a Bayesian stance distinct from that of Matthews'). Cumming (2012) goes into great detail about what is involved in deploying classical confidence intervals while rejecting the standard null hypothesis-testing scenario that is found in many textbooks.

I am taking no sides in this debate but it needs to be pointed out that many introductory statistics textbooks make a hash of teaching null hypothesis testing by conflating the quite distinct approaches of Neyman and Pearson and those of Fisher. Although this is done for reasons of simplification, the result is a hodge-podge that satisfies no one and has the inevitable result of rendering null hypothesis testing conceptually incoherent. It perhaps should be mentioned that neither Neyman nor Fisher liked the approach of the other, though "liked" may be too tame a term, especially when applied to Fisher.* In addition, there is research that shows that confidence intervals are hardly reported at all in most psychological research (Hoekstra et al. 2006).

Confidence intervals are, thus, not the neutral statistical procedure they might appear to be. They are one of the procedures at the interface in an internal classical statistics dispute between those who advocate null hypothesis testing and those who reject this approach. And for those who reject null hypothesis testing, what is available to them in the classical arsenal is statistical power, which everyone should use anyway, effect sizes, meta-analysis, and confidence intervals. It turns out that confidence intervals are more controversial than one might have thought and in ways other than those indicated by Gorard. This may help explain why Gorard concentrates on classical Cis in justifying the creation of his alternative, the trustworthiness matrix.

Gorard, however, goes on further to say, about confidence intervals, a statement whose import, it seems, is intended to be negative:

> Their true definition is an ideal, and it is recursive (involving itself) and reversed in logic (modus tollens). (p. 48)

In his earlier paper, Gorard additionally mentions that the use of modus tollens[8] is incompatible with what people use in everyday life, which is presumably modus ponens, and can thereby lead to confusion in properly interpreting CIs. Let me deal with these issues in order.

What Gorard means by "true definition" is unclear. I am not certain what he means by it. But the first thing to say is that no definition in the modern sense does this. The standard theory of definitions sets out what a term or phrase means within a given context, and its utility is its usefulness. That is all. Take the definition of "parallel line" as found in three contemporary geometric theories. In each of them, parallel lines are considered to be ideal in the sense that no empirical parallel line is "truly parallel" in the sense of each geometry. But the different parallel lines defined within or by the theories in which they are enclosed have each been found to be useful in its domain of application, which is all that is required of a definition. A theory, on the other hand, in addition to being useful, must also be true, in the sense of having a model (in the logician's sense). If this is taken to be a criticism of confidence intervals, it falls short of the mark.

That the definition of "confidence interval" is recursive is mentioned in a context that indicates that one should take this as a negative attribute of such a definition. This is a mistake. There is nothing inherently wrong in defining a term recursively, unless you make a mistake doing so, and lots of terms are so defined. One example is arithmetic addition. A recursive definition has a basis clause, a recursive (inductive) clause, and a closure clause. While there are a number of ways to define an expression, and though Gorard does not say so, it appears that his preference lies with explicit definitions, which are non-recursive. Whether one defines a term or phrase recursively or explicitly depends entirely upon the context. Gödel could not have constructed his famous incompleteness proof without the use of a recursive, indeed self-referential, definition.

This example of Gödel's proof makes me wonder whether Gorard has conflated recursion with self-reference. These two notions are neither identical with one another nor is their use inherently wrong. One just has to be careful. (I have lost track of the number of times I used

[8] **Modus ponens and modus tollens are rules of inference (with corresponding theorems). Their definitions are as follows. Modus ponens: If A implies B is true and A is true, then one can conclude that B is true. Modus tollens: If A implies B is true and not-B is true (B is false), then one can conclude that not-A is true (or A is false). These rules do not conform to ordinary English. They conform to the definition of the so-called material conditional, the sense of "implies" used here, which is the notion of implication used in virtually all of mathematics. There is a fallacious mode of inference known as the fallacy of affirming the consequent. It goes like this: A implies B is true. B is true. Therefore, A is true. This is a fallacious inference. It superficially resembles modus tollens in that you proceed from the consequent to the antecedent, but the procedure is invalid. We will come across this later.

to come across articles in math and logic journals that claim to have discovered a flaw in Gödel's proof, usually based on some discovered flaw in Gödel's use of self-reference.)

When we come to the conception of modus tollens where Gorard contends that the defining of confidence intervals is "reversed in logic", I have to confess that I have no idea what Gorard means. Modus tollens is a perfectly valid logical procedure. But his use of the term "reversed" makes me think that possibly Gorard has in mind the fallacious procedure of affirming the consequent, which can be viewed as a fallacious use of modus tollens, and that such a use is tied up with defining the concept recursively. This is so far from being right that it can not even be said to be wrong.

II. The Trustworthiness Matrix

I find the work by Gorard and his colleagues on their Trustworthiness Matrix inherently interesting; and it eerily resembles work done by myself and Funtowicz and Ravetz in the mid-eighties through the early nineties, though our applications are quite different.

While the matrix offers itself as a useful tool[9], Gorard makes some errors of interpretation and one of constriction in his discussion of the matrix.

As we have seen, the justification for Gorard and his colleagues developing the Trustworthiness Matrix seems to be the problems inherent in and around the use of confidence intervals. However, as with the matrices constructed by Ravetz, Funtowicz and myself in the past, the matrix itself is independent of the conflict that seems to encumber the confidence interval debate.

When we come to Gorard's Trustworthiness Matrix, we come to the core of the paper, although Gorard views the confidence interval discussion essential for motivating the construction of the matrix. In essence, the matrix provides a quasi-quantitative index of what Gorard calls trustworthiness, which is a type of quality assessment. Constructing such a quality assessment matrix is a difficult and worthwhile endeavor and Gorard and his colleagues have created a tool that researchers should find useful and insightful. However, like the context in which the matrix was developed (discussed above), some of the things that Gorard says about his matrix and its use and interpretation are essentially misleading or wrong. These comments do not take away from the utility of the matrix itself, however, as these errors can be fixed.

[9] I think I would consider a tool or method to be useful if it gave me a way to answer a question. In the case at hand, the question is: how do we assess the trustworthiness of a research project and/or its results? The matrix is the tool, or method, whereby we can find an answer. The complexity of the matrix shows that the answer to the question of trustworthiness is not an easy or completely straightforward one.

In constructing the matrix, what Gorard has done is to select a number of factors (the column headings) and associated with each of them an ordinal set of descriptive characteristics, what I will call descriptors. There are six column headings (factors) and five ordinally ordered descriptors for each factor. When characterizing the trustworthiness of a research project or program, a descriptor from each column is selected and the resulting set of descriptors itself placed in an ordered sequence, such as in this fictitious example, <4,3,2,2,1,0>. Thus, you have ordinal descriptors used to form an ordinal sequence. What you have in effect is an ordered 6-tuple. In this procedure, it is important for the column headings, the trustworthiness characteristics, or factors, to be logically independent. This is to prevent logical dependencies within the matrix itself from vitiating its interpretation. In addition, the stars Gorard uses can be eliminated, as they add nothing and, in fact, are intrusive.

In setting out the six factors characterizing the matrix, I think Gorard makes an unfortunate use of terminology. One of the factors is “Scale”. It is most easily seen if I reproduce Gorard’s matrix here (from Radstats 110, p.54).

Table 1: A ‘sieve’ to assist in the estimation of trustworthiness

Design	Scale	Dropout	Outcomes	Fidelity	Validity	Rating
Fair design for comparison	Large number of cases per comparison group	Minimal attrition, no evidence of impact on findings	Standardised pre-specified independent outcome	Clear intervention, uniform delivery	No evidence of diffusion or other threat	4★
Balanced comparison	Medium number of cases per comparison group	Some initial imbalance or attrition	Pre-specified outcome, not standardised or not independent	Clear intervention, unintended variation in delivery	Little evidence of diffusion or other threat	3★
Matched comparison	Small number of cases per comparison group	Initial imbalance or moderate attrition	Not pre-specified but valid outcome	Unclear intervention, with variation in delivery	Evidence of experimenter effect, diffusion or other threat	2★
Comparison with poor or no equivalence	Very small number of cases per comparison group	Substantial imbalance and/or high attrition	Outcome with issues of validity or appropriateness	Poorly specified intervention	Strong indication of experimenter effect, diffusion or other threat	1★
No report of comparator	A trivial scale of study, or N unclear	Attrition not reported or too high for any comparison	Too many outcomes, weak measures, or poor reliability	No clearly defined intervention	No consideration of threats to validity	0

Here we can see that the second factor is “Scale”. There are two reasons I think this an unfortunate choice of terminology. One is that the descriptors refer to the number of cases, or sample sizes. While it is true that “scale” is colloquially used in the way Gorard indicates in the matrix, a more appropriate term for this factor in my view would be “Sample”. Since the matrix itself produces a scale, an ordinal one, with “Scale” as one of the factors of the matrix itself, we are faced with an inherent equivocation. With that small alteration, the matrix seems well constructed.

When Gorard describes how the matrix is to be interpreted, however, he shows that the desideratum of logical independence of the fundamental trustworthiness factors is violated. In applying the trustworthiness matrix to any piece of research, what you obtain, as shown above, is an ordered set of numbers (an ordered 6-tuple), one for each descriptor (cell entry) of a fundamental factor (column heading). From bottom to top, these are scaled ordinally from 0 to 4. However, instead of independently selecting each descriptor for indexing purposes, Gorard asserts that in scoring the research, in going left to right, one must also go from top to bottom. That is, no descriptor score to the right of a given descriptor score can be higher than the one immediately to its left. This cannot be right, for the simple reason that the fundamental factors, the column headings, are or should be logically independent of one another. Gorard’s restriction means that one could never obtain a score like this: <4,4,0,1,2,3>. A perusal of the descriptors shows that this restriction hampers the effectiveness of the utility of the matrix and violates the inherent logical structure of the trustworthiness matrix itself.

There is another problem in Gorard’s scoring procedure. In simplifying his scoring procedure, Gorard suggests that the trustworthiness score given the research be the lowest score of the ordered 6-tuple, much like a chain is weakest at its weakest link. While there is nothing technically wrong with this, I would not recommend it as a general strategy, as such a strategy can misrepresent a research result’s trustworthiness. In going down this route, I would suggest a strategy that myself (Brownstein 1987) and Funtowicz and Ravetz (1991) developed for this purpose, which was to embed the matrix’s 6-tuple within another matrix, say a reliability matrix, whose purpose was to convert the 6-tuple into a single score or index. Problems encountered in such conversions are exacerbated if any of the individual descriptors are weighted in any way.

Although there is nothing preventing you from converting a trustworthiness score qua ordered 6-tuple to a single index, you lose information in so doing. In presenting the entire 6-tuple as the trustworthiness index, the analyst can then produce a more complicated and nuanced narrative than s/he might be able to do otherwise. Hence, we argued that the context determine which of these strategies in any individual case ought to be pursued.

While these criticisms do not in my view undermine what Gorard and his colleagues are trying to achieve and, indeed, have achieved in respect of their Trustworthiness Matrix, the errors committed in the discussion of confidence intervals, a discussion which is intended to motivate the construction of the matrix, are serious ones, as are the errors that Gorard commits in his discussion of the interpretation of the Matrix itself. Having engaged in this kind of analysis myself, I feel it is essential that no logical errors are committed nor that any mistakes be perpetrated in the interpretation of such a matrix or in the justifications surrounding it, as this could easily vitiate any utility such a matrix might otherwise possess.

*A brief discussion of the history of the dispute between Fisher and Neyman (sometimes including Pearson) can be found in Erich Lehmann, *Fisher, Neyman, and the Creation of Classical Statistics* (Springer: 2011), but you need to know your basic statistics in order to understand the discussion, as Lehmann assumes a good deal of the reader.

References:

L. Brownstein, "Relevance of the Rationalist/Constructivist/Relativist Controversy for the 'Validation' of Scientific Knowledge-Claims", *Knowledge, Diffusion, Utilization* (September 1987)

L. Brownstein, Conversation with J. R. Ravetz (June 2014)

G. Cumming, *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. (Routledge: 2012)

S. O. Funtowicz and J. R. Ravetz, *Uncertainty and Quality in Science for Policy* (Kluwer: May 1991)

S. Gorard, "Confidence intervals, missing data and imputation: a salutary illustration". MS. (No date.)

R. Hoekstra, et al., "Probability as Certainty: Dichotomous Thinking and the Misuse of *p*-values", *Psychonomic Bulletin and Review* (2006)

E. T. Jaynes, "Confidence Intervals Vs Bayesian Intervals". In Hooker and Harper, eds., *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*. Vol II. (Reidel: 1976)

R. Matthews, Bayesian Critique of Statistics in Health: The Great Health Hoax. (1998) http://www2.isye.gatech.edu/~brani/isyebayes/bank/pvalue.pdf

Larry Brownstein. Email: zen143717@zen.co.uk