Title
Reply to Rodgers and Rowe (2002)

Permalink
https://escholarship.org/uc/item/9rc977dd

Journal
Psychological Review, 109(3)

Authors
Roberts, Seth
Pashler, Harold

Publication Date
2002

Peer reviewed

Reply to Rodgers and Rowe (2002)

Seth Roberts and Harold Pashler

University of California, Berkeley and University of California, San Diego

Running head: Reply to Rodgers and Rowe

Please send proofs to

Seth Roberts

Department of Psychology

University of California

Berkeley CA 94720-1650

email roberts@berkeley.edu

phone 510.418.7753

fax   510.642.5293

Abstract

That a theory fits data is meaningful only if it was plausible that the theory would not fit.

Roberts and Pashler (2000) knew of no enduring theories initially supported by good fits

alone (good fits, that is, where it was not clear that the theory could have plausibly failed to

fit). Rodgers and Rowe (in press) claim to provide six examples. Their three non-

psychological examples (Kepler et al.) are instances of what we consider good practice:

How the theory constrained outcomes was clear, so it was easy to see that data might

plausibly have contradicted it. Their three psychological examples are flawed in various

ways. It remains possible that no examples exist.

Reply to Rodgers and Rowe (2002)

Were we to sit down at a table with Rodgers and Rowe, we like to think that the

four of us could eventually agree on many things–for instance, that there are important

differences between Mendel's support for his theory and goodness-of-fit evidence for

mathematical learning theories. We might not agree, however, on a point that Roberts and

Pashler (2000) did not make clear enough: <u>When used to evaluate complex psychological</u>

<u>theories, goodness-of-fit tests have been too easy to pass</u>. The theories are too flexible, the

data too variable, and contradictory outcomes too implausible. We fear we did not state

this bluntly enough.

In many cities, taxi drivers must pass a written test. Current practice in major

journals, such as <u>Psychological Review</u>–which often publish theories supported by no more

than good fits--resembles allowing a driver to pass the test if he can write his name.

"Theory testing should begin (but not end) with good empirical fits" is the title of Rodgers

and Rowe (in press). We agree. Taxi-driver tests should begin (but not end) with the test

taker writing his name on the test. "Unless a research can demonstrate that a theory

matches the data to begin with, there's little point in considering how it constrains possible

outcomes, how variable the data are, or whether results are surprising" (Rodgers & Rowe,

2002, p. 602). We agree. Unless a test-taker can write his name, there is little point in

considering whether he has more difficult-to-acquire skills. Rodgers and Rowe (2002, p.

600) note "the many uses of goodness-of-fit". We agree–it does have many legitimate uses,

just as it is often helpful to learn if a person can write his name.

The proof of a test is the competence of those who pass it. Many people can write their name who would not be good taxi drivers. What about the psychological theories that have passed goodness-of-fit tests and nothing else (tests without the additional information that we argue is necessary)? Many such theories have been published. How have they fared? Roberts and Pashler (2000) emphasized that "there seem to be no examples of a theory supported mainly by good fits that has led to demonstrable progress" (p. 358)–progress, that is, measured by a better metric than goodness of fit. This is a important difference, we hope everyone can agree, between our critique and criticisms of null hypothesis significance testing (NHST). Not even the most severe critics of NHST have claimed that it has never been associated with work of lasting importance (a claim which is obviously false). Even good tools can be misused. In Roberts and Pashler (2000), we were not saying that the use of good fits as evidence for complex theories had been misused. We were saying–perhaps too politely, perhaps not–that the practice is rotten to the core, in the sense that a taxi-driver test that can be passed by writing your name is rotten to the core.

Rodgers and Rowe (2002) disagree, of course. This is why their examples are the most important part of their argument. If current practice is reasonable–if passing a goodness-of-fit test <u>by itself</u> implies that a theory should be taken seriously by other scientists--then many theories that initially passed such tests (without the additional information we argue is necessary) should have turned out to have lasting importance. It is

telling that Rodgers and Rowe's first three examples (Kepler, Mendel, Watson & Crick)

are outside psychology–an indication, we suppose, that examples <u>within</u> psychology are

hard to find.

In Roberts and Pashler (2000), we tried to contrast what we considered bad

practice (testing goodness of fit) with what we considered good practice (testing

predictions). From Rodgers and Rowe (2002) we see that we might have been clearer. We

presented the two as disjoint (non-overlapping), but the real relation is that testing

predictions (i.e., meaningful tests) is a subset of testing goodness of fit. Here is what we

should have said: <u>Consistency of theory and data is meaningful only if inconsistency was</u>

<u>plausible</u>. With simple theories–e.g., theories that predict straight lines–that the theory

might not have fit may be obvious, but with more complex theories it is usually not

obvious, and therefore needs to be made explicit. However, authors of complex theories

routinely omit the necessary information. To show that a theory could have plausibly failed

to fit requires three steps: (a) Determine a way in which the theory constrains possible

outcomes. Any prediction can be described in terms of a <u>forbidden zone</u> (Roberts &

Sternberg, 1993, p. 638) consisting of outcomes the theory cannot generate no matter

what parameter values are used. For example, if the prediction is that a certain value

computed from the results will be more than zero, then the forbidden zone is zero and

less. (b) Shrink the forbidden zone to allow for the variability of the data. If the initial

forbidden zone–not taking variability into account–is zero and less, and the variability in

the data (e.g., between-subject variation) give a 95% confidence interval of length 10 for

the value, then (assuming symmetric confidence intervals) the forbidden region shrinks to -

5 and less, because an observed value of (say) -3 would be compatible with the theory. (c)

Determine if any plausible outcomes lie in the shrunken forbidden zone. For example, do

plausible alternative theories generate outcomes in that zone? Only if the forbidden zone

contained plausible outcomes should consistency of theory and data impress us.

   We might have used Rodgers and Rowe's (2002) non-psychological examples as

instances of good practice. Kepler, Mendel, and Watson and Crick made explicit how their

theories constrained the data. Kepler derived Kepler's Laws, which are constraints. Mendel

pointed out that his theory predicted certain ratios of phenotypes. Watson and Crick noted

their theory predicted Chargaff's Rules. Because these constraints were made explicit (Step

1 of the three-step process described above), it was possible to take variability (Step 2) and

plausibility (Step 3) into account. In each case--Kepler, Mendel, Watson and Crick–readers

of the initial work could see that the theory narrowly constrained outcomes in cases where

a much wider range of outcomes were plausible. Readers could see, in other words, that

the theory could have easily failed the test. This is what many modelers do not make clear.

   The first of Rodgers and Rowe's (2002) three psychological examples is "the

Weber-Fechner psychophysical function" (p. 601). We are unsure what theory this refers

to; functions are not theories. Fechner (1860/1912) proposed that sensation grows as the

logarithm of stimulus strength but this was apparently a definition rather than a theory,

because sensation strength was not independently defined. Rodgers and Rowe's reference here (Falmagne, 1974) contains several theoretical ideas but as far as we know none of them has been influential.

Their second psychological example is Shepard's (1987) theory of stimulus generalization, built to explain an empirical "law" (Shepard, 1987, p. 1317) that Shepard noticed. We are less sure than Rodgers and Rowe that this theory has lasting value, but the more important point is that Shepard made explicit how his theory constrains the data (it implies the empirical law). The data are given in enough detail to take variability into account, and a careful reader can see that plausible outcomes would have contradicted the theory.

Rodgers and Rowe's final psychological example is Tversky and Kahneman's prospect theory. The curve-fitting to which Rodgers and Rowe refer (Figures 1 and 2 in Tversky and Kahneman, 1992) involves an equation not derived from the theory. The theory was supported by not by the good fit of this equation but by more general features of the data. Equation fitting has never been used to support prospect theory, as far as we know (Kahneman & Tversky, 2000).

Rodgers and Rowe (2002) do readers a valuable service by providing the six examples. None of their examples, we believe, supports their case. If any examples exist of what we could not find–a theory initially supported only by goodness of fit (without the necessary additional information) that turns out to be important–they are plainly far and

few between.

Rodgers and Rowe (2002) also defend their own work (Rodgers & Rowe, 1993), which Roberts and Pashler (2000) mentioned briefly. Because the practice we criticized is so common, it is a little unfair to spend much time on any one instance. Rodgers and Rowe (1993) were more sophisticated than most modelers. They understood the problem, or came close to understanding it: "We were concerned that the model we built to explain prevalence curves for adolescent sexual development might . . . explain any developmental process" (Rodgers & Rowe, 2002, p. 601). However, they did not correctly solve the problem, at least in our judgement, and why they failed is worth pointing out so that others can avoid their mistakes. "We fit a two-gender sexual development model to data for smoking onset and to several other artificial datasets. . . In each case the model was rejected" (Rodgers & Rowe, 2002, p. 601). One mistake is that they failed to consider variability. It must be shown that the theory could not have fit outcomes of <u>the procedure</u> <u>that produced the actual results</u> (including sample sizes). The smoking onset data came from the same survey as the sexual development data, but the sample sizes may have been different (due to non-responses). Maybe smoking questions were answered more often than sexual questions; a larger sample size makes it easier to reject the theory. The sample sizes of the artificial data sets are not stated. Second, and perhaps more importantly, they did not consider plausibility. Whether the smoking data and the artificial data were plausible outcomes of the sexual part of the survey is unclear. Rodgers and Rowe do not

present the smoking data and the artificial data, so the reader cannot decide. Their third

mistake is more subtle. To show that the theory cannot fit <u>one particular outcome</u> (or a

small number of outcomes) is not enough. The probability of any one outcome is usually

very low. To show that one particular outcome would falsify the theory is only to show

that one very unlikely event would do so. The data used by Rodgers and Rowe (1993) can

be described as a set of 96 integers, the number of students who responded "yes" in each of

the 96 cells of the data table. The sample size for each cell was about 50, so there were

about 51^{96} possible outcomes. (A sample size of <u>n</u> allows <u>n</u> + 1 possible outcomes. For

instance, with a sample size of 2, there are 3 possible outcomes: 0, 1, and 2 "yes" answers.)

Any one was very unlikely, of course. To show the theory might plausibly failed, it must be

shown that the theory cannot fit a <u>range</u> of plausible outcomes, not just 1 or 10 or 1000.

Rodgers and Rowe's (2002) discussions of aspects of variance (p. 600) and

plausibility (p. 600) are beside the point. To show that the good fits of Rodgers and Rowe

(1993) support the theory requires showing it was plausible the theory might not have

fit–nothing more, nothing less. The discussion of irrelevant issues is probably our fault:

We failed to make clear that the three issues we raised in Roberts and Pashler (2000)–the

need to determine what the theory cannot fit, to consider variability, and to consider

plausibility–were three parts of one task: showing that the theory might plausibly not have

fit.

Roberts and Pashler (2000) pointed out that typical uses of good fits to support

theories have lacked crucial information. This meant this type of evidence (good fit) <u>could</u>

be abused–used to claim more support for a theory than it deserves. We also pointed out

that this type of evidence had not been the initial evidence for any enduring theory, as far as

we knew. This suggested that this type of evidence <u>had</u> been abused. In spite of ample

opportunity, Rodgers and Rowe (2002) have failed to come up with even one clear

counter-example. This suggests that the abuse has been extensive.

References

Falmagne, J. C. (1974). Foundations of Fechnerian psychophysics. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), Contemporary developments in mathematical psychology, Volume II: Measurement, psychophysics, and neural information processing (pp. 127-159).

Fechner, G. T. (1912). Elements of psychophysics: Sections VII and XIV. In B. Rand (Ed.), The classical psychologists (pp. 562-572). Boston: Houghton Mifflin. Originally published 1860.

Kahneman, D., & Tversky, A. (2000). Choices, values, and frames. Cambridge, UK: Cambridge University Press.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. Psychological Review, 107, 358-367.

Roberts, S., & Sternberg, S. (1993). The meaning of additive reaction-time effects: Tests of three alternatives. In D. E. Meyer & S. Kornblum (Eds.), Attention and Performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience--A Silver Jubilee (pp. 611-653). Cambridge, MA: MIT Press.

Rodgers, J. L., & Rowe, D. C. (1993). Social contagion and adolescent sexual behavior: A developmental EMOSA model. Psychological Review, 100, 479-510.

Rodgers, J. L., & Rowe, D. C. (2002). Theory testing should begin (but not end) with good empirical fits. Psychological Review, 109, 599-603.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological

science. Science, 237, 1317-1323.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative

representation of uncertainty. Journal of Risk and Uncertainty, 5, 297-323.

Author Note

Seth Roberts, Department of Psychology; Harold Pashler, Department of

Psychology.

Correspondence concerning this article should be addressed to Seth Roberts,

Department of Psychology, University of California, Berkeley, California 94720-1650 or

roberts@socrates.berkeley.edu.