Theoretical knowledge and behavioral science

Frank Zenker, Bogazici University, Istanbul, Turkey (fzenker@gmail.com)

Long abstract: In behavioral science (e.g., social psychology) as elsewhere, a transparent evaluation of the scientific relevance of an observed or predicted effect should distinguish between the size of effect (*statistical aspect*), the ability of a theoretical construct to predict the effect (*theoretical aspect*), and the utility that is associated with the effect (*practical aspect*). Although often conflated, the practical aspect is independent of the theoretical and statistical aspect. On formal grounds, moreover, the 'smallest effect of interest' (SEOI) that is needed for a prediction is much larger than the SEOI for an explanation. Since recent meta-meta-analyses suggest that behavioral science publications report observed effects that are either small and homogenous or large(r) and heterogeneous, the SEOI that an empirically adequate theoretical construct would predict thus exceeds what can typically be observed. As this massively impairs theory development in behavioral science (partially explaining why such constructs remain rare), behavioral science broadly lacks theoretical knowledge today: some 50 years of experimental/observational research have so far failed to result in empirically adequate theories that allow to explain and predict human behavior, and intervene on it (e.g., in public policy making).

In explaining this "state-of-the-art" and discussing possible remedies, I raise three related issues. First, behavioral science research today is typically data-driven rather than theoretical (i.e., inductive rather than deductive). But on pains of running in Hume's problem of induction, not only are theories needed to explain the regularity that is presupposed by an inductive projection of past observations into the future; the content of an empirical theory must also exceed the observations it subsumes. The focus on induction thus suggests a lack of understanding among behavioral scientists what theoretical knowledge is, and why it is important.

Second, in individual studies and across replication studies, experimentally observed behavioral responses typically translate into homogenous yet small effects. Such effects, however, are overlain by the standard measurement error, thus becoming quasi-unobservable effects. Where observed behavioral responses are medium-sized (or larger), by contrast, they are typically heterogeneous (i.e., entail large observed variance), and so are unclear, or diffuse. Consequently, the point-effect that an empirically adequate theoretical construct would predict is often unknown, partly explaining why theoretical knowledge in behavioral science cannot easily arise.

Third, the best statistical inference strategies are regularly applied mindlessly; particularly the application of a Bayesian hypothesis support-threshold tends to ignore the minimum sample size. This is partly explained by a strong focus on conducting studies that are aimed at "discoveries" (by testing against chance), while avoiding a theoretical construct that is more informative than a directional hypothesis.

I present three related remedies. First, coordinated efforts among many labs allows studying overall fewer effects (given constant resources), yet would generate sufficiently large samples (needed to reduce the error-rates) and thus allow for improved effect size estimates. This entails studying overall fewer effects than is currently the

case, yet these fewer effects could be studied "much harder" than is the case today. The major question here is how to reorganize behavioral science such that cooperative research is properly incentivized.

Second, the large observed variance can be fruitfully addresses by specifying, and transparently reporting, assumptions that underlie a study's measurement error, which at the same time amounts to engaging in theory construction research. It is a contingent fact that behavioral science publications normally employ standardized effect size measures, that quantify the observed mean difference, $m_1 - m_0$, relative to the observed standard deviation, s. A prime example is Cohen's d-measure: $d = (m_1 - m_0)/s$, or its transformation into the correlation measure $r = d/\sqrt{d^2} + 4$. Standardized effect size measures are commonly used in meta-analytical research to quantify the observed $m_1 - m_0$ across object-level studies that use different measurement-scales, or in theory-construction research to point-specify $m_1 - m_0$ as a theoretically predicted parameter. Since standardization conceptually relates to the quality of measurement, the observed $m_1 - m_0$ can be fully interpreted only relative to the error-theory that determines s. This error-theory, however, must typically be chosen freely, because a theoretically motivated measurement-scale is normally unavailable. Using hypothetical but realistic data from educational research, it can be shown that differentially sophisticated error-theories let the observed $m_1 - m_0$ vary massively (given identical mean differences as observations), because the amount of effect results from the mean difference, the observed standard deviation, the dependent variable's reliability, the measurement error, the quasi-experimental setting's quality, and from how this setting is standardized. This lets the common praxis of publishing standardized effect sizes "nakedly"—without a transparent error-theory—appear problematic, because this praxis undermines the goals of a cumulative science of human behavior.

Third, and finally, the large observed variance of behavioral responses can be reduced in a post-hoc manner by projecting its statistical representation (i.e., a probability density) into anthropometrically valid measurement-scales. Information that is seemingly "lost" will now be compensated by gaining a clearer view on the observed effect. Tongue-in-cheek, we call this a "quantum-measurement" approach. Crucially, the approach assumes that the human ability to differentiate psychological states is limited—namely to a finite set of quanta—rather than allowing for the continuous differentiations that a statistical treatment of behavioral responses does misleadingly suggest. We exemplify this by means of simulation results.