HHS Public Access

Author manuscript

Psychol Inq. Author manuscript; available in PMC 2016 March 24.

Published in final edited form as:

Psychol Inq. 2015 ; 26(3): 286–292. doi:10.1080/1047840X.2015.1064066.

THE ART OF SMART SCIENCE: WEAVING THEORY AND RISKY STUDY DESIGN INTO PSYCHOPATHOLOGY RESEARCH AND RDOC

Uma Vaidyanathan[a], **Scott I. Vrieze**[b], and **William G. Iacono**[a]

[a]University of Minnesota

[b]University of Colorado Boulder

Yes, the art. We are grateful for the opportunity to respond to commentaries on our target article. Just like the commentators, the commentaries were excellent, expanding on the target article in important ways while simultaneously taking us to task in others. On one thing each commentator agreed: there is a lot of room to improve how we conduct research to advance knowledge. However, there was less agreement on how best to do that, reflecting the complex and interrelated nature of experimental design, statistics, multi-method inference, and causal inference, which were the focus of our target article. The commentaries all expounded on these issues in important ways and many took the topic further, emphasizing the importance of clinical application, tradition, and health policy.

Special Issue Recap and Overview of Response

We have summarized in Table 1 the chief arguments for our collective main theses and the solutions proposed to move the field forward. We delve further into these themes in this article, cutting across them to emphasize commonalities and grouping topics in our discussion to tease out philosophical and policy perspectives. It was unfortunately not feasible to cover in our response all the themes that were raised, in no small part due to the extensive experience and expertise represented by our reviewers. Our goal in this response is to weave these thoughtful perspectives into a bigger picture and propose a way forward for mental health science.

Theme 1: Philosophy, Data, Theory, and the Puzzle of How Best to Integrate Them

Our target article dealt with issues relating to study design and evaluation, as well as what constitutes strong evidence for and against particular etiological theories. Most of the responses we received were likewise focused on these themes.

Author Information:

Uma Vaidyanathan: Department of Psychology, University of Minnesota, N218 Elliot Hall, 75 East River Road, Minneapolis, MN 55455. vaidy017umn.edu

Scott I. Vrieze: Department of Psychology & Neuroscience, Institute for Behavioral Genetics, University of Colorado Boulder, 1480 30th Street, Boulder, CO 80303. scott.vriezecolorado.edu

William G. Iacono: Department of Psychology, University of Minnesota, N218 Elliot Hall, 75 East River Road, Minneapolis, MN 55455. wiaconoumn.edu

I. Validity and Measurement

A number of commentators, including Lilienfeld and Pinto (this issue), and Markon (this issue), explicated the role of various types of validity, and emphasized the importance of measurement issues involved in study design. Lilienfeld and Pinto (this issue) agreed with our arguments that the measures and methods we use are imperfect indicators of the phenomena we are attempting to study. We wholeheartedly concur that it is important to understand measurement properties for a construct of interest, and whether the methods scientists use provide a stringent means to test substantive hypotheses, methods that go beyond mere focus on a t-test or ANOVA result (as Miller and Yee (this issue) noted). Moreover, as emphasized by Lilienfeld and Pinto (this issue), if the construct we are measuring is a robust one and has some real world meaning, results from different studies should converge, regardless of the statistic, questionnaire, or methodology used.

Markon (this issue), on the other hand, emphasized more the roles of statistics, ontology and parsimony in shaping scientific discourse and inference. We agree there is great value in quantification and precise measurement; our point was that focusing on these qualities in the absence of a thorough understanding of their limits, a good research design, and a strong theory will not yield substantial etiological insight. An example of this comes from a recent meta-analysis of almost every twin study undertaken from 1958 – 2012 (Polderman et al., 2015). This meta-analysis investigated the heritability of almost every medical and psychiatric condition from twin studies using 14.5 million twin pairs. Clearly, this undertaking represents an important integration and summary of decades of twin research that will be useful as a reference for years to come. The results suggested that most traits fit an additive genetic model with about 49% of variation attributable to additive genetic factors on average. Some exceptions were noted for psychiatric conditions like conduct disorder and recurrent major depression where the pattern of twin correlation suggested shared environment or non-additive genetic factors. However, as the authors of the paper pointed out, this study could not pinpoint the reason for the "missing heritability" indicated from genome wide association studies showing that observed molecular genetic variation cannot account fully for the heritability estimates derived from twin studies, or additional sources of genetic variation such as those derived from non-additive genetic effects. In fact, they opine that in the latter case, we need data "…for example, from large population samples with extensive phenotypic and DNA sequence information, detailed measures of environmental exposures and larger pedigrees including non-twin relationships" (p. 7). In other words, despite the large sample size and sophisticated statistics, because there are no risky tests or hypotheses involved here, it is difficult to further our knowledge about the etiology of any of these disorders beyond what the twin correlation patterns tell us.

II. Research design versus statistical inference and how they affect causal inference

While we emphasized a risky test and a good research design as being of central importance, several commentators noted the need for large sample sizes and replicability as a fundamental issue. Ioannidis' (this issue) commentary is an exemplar of this perspective; Lilienfeld and Pinto (this issue) likewise noted the need for replicability and highlighted the importance of convergence of indicators. These are all additional important components of risky test taking because they all increase the likelihood of disconfirming or at least pointing

to the limitations of a theory. But what type of replication is most useful under what circumstances, and is it enough to just ask for large samples? How large is large enough – hundreds, thousands, tens of thousands, millions, or simply large enough to reflect accurate capture of a predicted effect size related to the question at hand? One way to answer this question (without specifying some arbitrary N) would be to require studies to provide a rationale (based on the state of prior knowledge) as to why a sample should be adequately powered to detect some predicted effect size (see Miller and Yee (this issue) for a similar point), to avoid publication of results from small studies that could be statistically significant based on chance alone (e.g., Button et al. (2013) found that the median statistical power in neuroscience studies is 21%), or are so underpowered that a null finding is uninformative. Genetic association studies of complex traits and diseases must be very large because the expected effect sizes are small, with "large" effects accounting for a fraction of a percent of variance for continuous traits, for example, whether those traits are questionnaire responses, EEG recordings, or a brain scan.

Our point is that while all design elements (e.g., experimental design, extremely large samples, replicability, etc.) work together to provide useful additional evidence regarding a theory's robustness and the limits of its applicability, none of these factors alone is sufficient to confirm or disprove a theory. Consider a simple example – every day millions of people observe the sun moving from east to west from different parts of the world; multiple repeated measurements are obtained from a very large sample, but as we all know, the sun does not revolve around the earth. In and of itself, a large sample does not help here – and actually leads to the incorrect conclusion in this case. Simply put, it is not risky enough a test on its own. And yet, the geocentric or Ptolemaic system was dominant for hundreds of years. It was only with the addition of other information that could not be confirmed with the naked eye and theoretical postulates which were later supported (e.g., elliptical orbits, phases of Venus, Jupiter's moons) that the test became riskier, the data failed to fit the geocentric theory, and the findings were deemed to fit the Copernican theory better. It was the combination of multiple elements of theory testing that falsified the geocentric theory.

III. How does one decide what is sufficient evidence for a theory?

Agrawal and Bogdan (this issue), Ioannidis (this issue), and Widiger, Crego, and Oltmanns (this issue) all noted that there are no objective standards to determine how to optimally interpret findings in such a way that technology, statistics, and multiple converging lines of evidence from different levels of analysis can be used to decide what constitutes sufficient empirical support. Because there is no specific numerical cutoff or entirely objective criterion, we should strive to develop and use research designs that put competing theories at risk. We contend (despite Markon's (this issue) or Ioannidis' (this issue) assertions) that almost any criterion used to select theories such as "good fit", "parsimony", "harm minimization", etc. all contain an element of subjectivity that renders them difficult to adopt in a universally accepted, objective manner. This is where the art of science applies – in positing theories that may sometimes go beyond what seems reasonable given existing knowledge (e.g., as in the example provided above about whether the geocentric or Copernican models reflected reality), while evaluating them using well thought out research designs that provide risky tests and narrow the number of interpretative possibilities.

Agrawal and Bogdan (this issue) present a compelling example of the successful application of a risky test in their paper where they review evidence from quasi-experimental research showing that early marijuana use results in higher risk for use of other substances later on. While the results they describe are consistent with a "gateway" model of causation, follow-up studies can test more directly the gateway interpretation, and rule out alternative interpretations. For example, it may be that deviant peers affect both early marijuana use and later hard substance use, a possibility that could be explored using twins concordant for early marijuana use who are discordant for deviant peer relationships. Unfortunately, such twin pairs are not typical, rendering difficult ascertainment of a sufficiently large sample and possibly leaving unanswered questions regarding the generalizability of results. However, our point is that, when possible, we should capitalize on and value the results of such studies and continually look for complementary ways to put the theory at further risk. In this case, for instance, we may rely on a longitudinal study of more readily ascertainable discordant siblings instead of twins, or evaluate in a purely observational sample the effect of naturalistic switching of peer groups among early marijuana users as an instrumental variable. Far more clever designs whether observational, quasi-experimental, or even experimental, are surely possible.

We also acknowledge Klein and Hajcak's (this issue) commentary in this regard. They not only expand upon our example of recurrent depression by providing a more comprehensive and in-depth investigation of the topic (differences in correlates of recurrent vs. single episode depression) by including more indices of neurobiology, self-report, diagnostic data, and functional outcomes in their response, they also include discussion of research using observational and quasi-experimental designs. Together these two commentaries (Agrawal & Bogdan, this issue; Klein & Hajcak, this issue) represent precisely the kind of integrative thinking and risky testing we intended our article to spur. In our opinion, these are examples of the path our field should take to make substantive gains in knowledge.

Theme 2. Implications for Science Policy and Incorporating the Human Element into Research

A second theme in many of the commentaries is to redesign the incentive system in science to better support the accumulation of knowledge rather than focusing on publishing alone as an endpoint. As scientists, we have a responsibility to ourselves, the research community, and the public at large to put honesty and accuracy at the forefront of our research and communicate results accordingly. However, as much as science is considered to be an objective profession, as we have repeatedly attempted to underscore throughout our discussion, certain fundamental concepts in our field – e.g., "parsimony", "harm", "clinical significance", "utility", etc. – are not quantifiable using a universally accepted metric. Neither do we operate in a purely scientific vacuum that is unhindered by concerns such as job security, funding, acceptance by peers, desire for fame and prestige, and so on. Our current system based on publications and peer review all but ignores such human elements, and focuses simply on outcomes such as number of publications, impact factors of journals that we publish in, amount of grant funding and so on. This set of external contingencies affects science in two ways – at the group level, in which fields like psychiatry embrace a

certain school of thought (e.g., currently neurobiology and genetics), and at the individual level (where scientists have pressures to produce results – any result, not necessarily accurate ones.).

I. Science as a political process

A couple of our commentators (Widiger, et al., this issue; Zachar, this issue) alluded to this issue, pointing out correctly that science is a competitive, political process. They noted that self-critical examination of pet theories by scientists is essential, and that competing viewpoints need to be acknowledged and discussed. This is perhaps especially true for major decisions that affect healthcare worldwide, such as the classification of personality disorders by the DSM 5 workgroup (as outlined by Widiger et al. (this issue)). We concur with our commentators that risky tests are a good way to arbitrate between competing models even in such political decisions. Our commentators also stressed that convergence between results from disparate methodologies and domains is a key component in building robust theories with explanatory potential. However, as we noted earlier, questions of utility can be distinct from questions of etiology. One may not need a rigorous experimental design to test whether people get better after receiving psychotherapy, for example, although whether the psychotherapy caused the improvement is a question that requires quasi-experimental and experimental research. Similarly, knowledge of a mechanistic causal relationship between variants in nicotinic receptor gene and increased cigarette smoking may have zero impact on clinical treatment if the causal chain explains only a tiny fraction of risk for smoking. That said, knowledge of etiology in both cases is useful when making informed decisions about how to allocate resources for treatment. If psychotherapy has no causal relationship with improved outcome, then one could imagine replacing psychotherapy with a less expensive alternative, with no detriment to the patient. If the all genetic effects within a candidate gene(s) are known to be very near zero, then targeting that system for therapeutic development may not be appropriate.

Political decisions are, well, political, but certainly can be informed by scientific understanding. In turn, we contend that scientific understanding is accelerated when scientists undertake research programs with study designs that permit risky tests of etiological theories from multiple angles – whether those risky tests involve large samples, multiple methodologies, (quasi-)experimental designs, longitudinal data, or some variation and/or combination of all the above.

II. Scientists as faulty human beings

Aside from the political issues noted above, we think there is a broader issue of responsible science at the individual level. Currently, as most readers are aware, research involves obtaining resources to fund research, collecting data, performing some sort of study (e.g., in a lab setting, running a statistical model), getting a (statistically significant) result, writing up said result in a manuscript, submitting it for peer review and hopefully publishing it, and repeating the cycle.

The number of publications affects job security as well as future grant funding, regardless of their scientific quality, reproducibility, and actual contribution to the state of knowledge in a

particular field. One of our commentators (Ioannidis, 2011) has already shown very elegantly that there are more studies in the volumetric brain imaging literature with statistically significant results, than what would be expected from power calculations using sample sizes in those studies alone. Others (Fanelli, 2010, 2011) have similarly noted that the number of positive findings in published scientific papers across all fields has increased by 22% from 1990 to 2007; this increase was especially marked for the social sciences in which "…the odds of reporting a positive result were around 5 times higher among papers in the disciplines of Psychology and Psychiatry and Economics and Business compared to Space Science" (p. e10068). This hyper-emphasis on the number of publications with positive findings (and grants) is occurring in the context of decreasing funding for research, as noted by the head of NIH, Francis Collins (Szabo, 2014), leading to decreasing numbers of academic positions for young investigators (Harris, 2014).

An additional problem that has been gaining more attention in recent years is that of researchers falsifying or more commonly, being unknowingly careless with their data and analysis. Several high profile recent cases include that of Diederik Stapel (Levelt Committee, Noort Committee, & Drenth Committee, 2015), Marc Hauser (Department of Health and Human Services, 2012), Andrew Wakefield (Dominus, 2011), to name a few. The true incidence of falsification is unknown, though estimates of irreproducible research from fields such as preclinical research exceed 50% and cost about $28 billion per year (Freedman, Cockburn, & Simcoe, 2015). Likewise, the frequency of retractions has been found to be strongly correlated with the impact factor of a journal (Fang & Casadevall, 2011). While speculative, results such as these suggest that the pressure to publish positive findings might motivate some to engage in questionable data analytic practices, knowingly or unknowingly. An uber-competitive system that emphasizes number of publications and promotes a winner-takes-all approach (all funding, all jobs, all big ideas, all credit), does little to foster responsible research, scientific practice and, by corollary, accumulation of knowledge. It is impossible to prevent any researcher from ever falsifying data or engaging in questionable research practices. However, it is possible to modify the current incentive system to overcome or mitigate some of these challenges.

How can we design such a system? As a first step, we can stop relying exclusively on the number of publications, or publications in high-tier journals as a primary metric of good research. What would be a good alternative? Perhaps whether an investigator enters into multi-site, team-based collaborations, or perhaps the number of times they share their data for replication with other investigators, or even the number of times they attempt to undertake replications of their or others' findings. Note that such metrics are independent of whether some researcher gets a positive or negative result. In other words, we reward acts like collaboration and data sharing rather than exclusively focusing on the outcome, which would result in larger datasets for analyses, greater transparency in procedures used, and multiple theoretical perspectives to analyzing data. The 1000 Genomes Project is a great example of how the power of combined datasets and public data release can lead to greater knowledge about the topic they are focused on (The 1000 Genomes Project Consortium, 2012). 1000 Genomes was especially powerful because all raw data is public – anyone can download it. While more restrictive than 1000 Genomes, several data repositories have been formed in recent years such as the Database of Genotypes and Phenotypes (dbGAP; http://

www.ncbi.nlm.nih.gov/gap), the National Database of Autism Research (NDAR; https://ndar.nih.gov/), and Research Domain Criteria database (RDoCdb; http://rdocdb.nimh.nih.gov/). What is needed at this point are incentives to submit to, contribute to, and utilize such databases. In this regard, it is encouraging that NIH and other organizations have set out guidelines and started various initiatives to encourage replicable work and replication amongst researchers (Bobrow, 2015; Collins & Tabak, 2014; NIH, 2015).

Second, as Miller and Yee (this issue) suggested, we could build good theory testing as a peer review criterion for a grant proposal or manuscript submitted for publication. For example, NIH uses a peer review system to evaluate grants where each reviewer is asked to take into account the following five criteria: significance, suitability of investigators, innovation, research approach, and suitability of the environment. Ioannidis, in the recent past, has been a vocal critic of this evaluative system (see Nicholson & Ioannidis, 2012 and series of responses from the Office of Portfolio Analysis at NIH and others in *Nature*), noting that NIH rarely ever funds truly innovative research. Perhaps the problem here in part is relying on a subjective criterion such as innovation, which as Miller and Yee (this issue) posited, may involve little more than application of novel technology. On the other hand, if the grant review process assigned greater value to a research design that provided a risky test of an influential theory, then both investigators and peer reviewers would be more likely to recognize the merit of research that proposes a risky test involving a well conceptualized theory.

Third, journals could prioritize publishing adequately powered and designed replications, and mirroring this, granting institutions and departments could also weigh heavily the value of replications in their grant review and/or tenure process. To some extent, we cannot blame investigators for not wanting to attempt replication studies if journal editors do not want to publish them, and research departments and institutions do not view replication as nearly as important as a faculty member's ability to establish an independent line of research. While one objection to encouraging a program of replication would be that it would slow down progress, we believe the opposite would happen instead: It would ensure that major studies are replicated by different scientists, thereby providing the solid foundation needed to justify subsequent investment of resources to build on the findings. In this context, it is encouraging to see replication ventures such as the *Reproducibility Project: Psychology* and *Many Labs* receive much positive publicity and editorials in conventional journals like *Nature* (Baker, 2015; Yong, 2013).

Conclusion

We have attempted to synthesize and draw out the common themes from among the varied perspectives provided by our commentators. We were pleased to see that all our commentators supported our general conclusion that the way research in psychology and psychiatry is currently conducted is not satisfactory; the way forward, though, was not as clear. In our rejoinder, we have proposed that the key is to design a system of science that rewards undertaking risky, collaborative, and replicable research, rather than focusing

merely on a particular methodology, or feature of research such as statistics, novel technologies, or large sample sizes.

We are not alone in such calls for reforming research in psychology. As mentioned earlier, Ioannidis, Fanelli, and their colleagues have been very active this field. Likewise, Brian Nosek and Yoav Bar-Anan (2012) have published in this same journal (see Psychological Inquiry Vol. 23(3) for target article and commentaries), arguing the need for a revamped system for scientific communication – especially one that is undergirded by openness and transparency at all levels, including the availability of data, the peer review process, and continuous post publication review. Nosek has a taken such calls one step further, and founded the Center for Open Science (COS), which attempts to foster exactly the kind of work he and his co-author outlined in their target article (Nosek & Bar-Anan, 2012). We are in complete agreement with their efforts and find commendable that he and his colleagues not only "talked the talk" but are "walking the walk"!

Another initiative that was mentioned quite often throughout the commentaries was the Research Domain Criteria (RDoC; Insel et al., 2010). Several commentators offered perspectives on RDoC and provided suggestions for its improvement. Regier (this issue) exhorted the critical need for systems such as RDoC while urging NIMH to rely not just on basic science research, but focus on clinical, epidemiological, and health services research as well. Similarly, Zachar (this issue) emphasized "convergence seeking is what RDoC should evolve into". Likewise, Kagan's (this issue) central point is also highly pertinent to RDoC: against the backdrop of biology and genetics, it is nevertheless the case that the environment profoundly impacts what is or is not perceived as a disorder. Miller and Yee (this issue), based on their communications with RDoC workgroup members, noted that this is indeed the case – that in actuality, RDoC is not reductionistic and that it does incorporate levels of analysis ranging from the biological to the psychological, as can be seen from the RDoC matrix. RDoC has the potential to improve research in ways that we and many of our commentators would likely endorse. In a recent blog post, the Director of NIMH, Thomas Insel, refers to RDoC as "convergent science" and as "bringing together many levels of analysis" (Insel, 2015). Bruce Cuthbert, the Director of the RDoC Unit at NIMH, has likewise noted that constructs included in the RDoC matrix had to be defined in terms of some behavioral or cognitive process, be linked to a neural circuit, and be relevant to psychopathology (Cuthbert, 2015), thus emphasizing the importance of a conceptual (if not quite theoretical) connection across research domains.

It is heartening to see that our field is starting to evolve in the various directions proposed by our commentators including replicability, large scale research, and encouraging convergence amongst various types of measures. We would still contend that neither addresses explicitly what we consider a linchpin of good research that holds all these elements in place – i.e., risky tests.

We started our target article with a quote, and would like to book-end our response with another one:

"After a certain high level of technical skill is achieved, science and art tend to coalesce in esthetics, plasticity, and form. The greatest scientists are always artists as well."

—Albert Einstein

References

Agrawal A, Bogdan R. Risky business: Pathways to progress in biologically informed studies of psychopathology. Psychological Inquiry. (this issue).

Baker M. First results from psychology's largest reproducibility test. Nature. 2015

Bobrow M. Funders must encourage scientists to share. Nature. 2015; 522(7555):129–129. [PubMed: 26062475]

Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafò MR. Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience. 2013; 14(5):365–376.

Collins FS, Tabak LA. NIH plans to enhance reproducibility. Nature. 2014; 505(7485):612–613. [PubMed: 24482835]

Cuthbert BN. Research Domain Criteria: toward future psychiatric nosologies. Dialogues in clinical neuroscience. 2015; 17(1):89. [PubMed: 25987867]

Department of Health and Human Services. Case Summary: Hauser, Marc. 2012. Retrieved from https://ori.hhs.gov/content/case-summary-hauser-marc

Dominus S. The crash and burn of an autism guru. 2011 from http://www.nytimes.com/2011/04/24/magazine/mag-24Autism-t.html?_r=0.

Fanelli D. "Positive" Results Increase Down the Hierarchy of the Sciences. PLoS ONE. 2010; 5(4):e10068. [PubMed: 20383332]

Fanelli D. Negative results are disappearing from most disciplines and countries. Scientometrics. 2011; 90(3):891–904.

Fang FC, Casadevall A. Retracted Science and the Retraction Index. Infection and Immunity. 2011; 79(10):3855–3859. [PubMed: 21825063]

Freedman LP, Cockburn IM, Simcoe TS. The Economics of Reproducibility in Preclinical Research. PLoS Biol. 2015; 13(6):e1002165. [PubMed: 26057340]

Harris R. Too few university jobs for America's young scientists. NPR. org. 2014

Insel, T. Crowdsourcing RDoC. 2015. Retrieved from http://www.nimh.nih.gov/about/director/2015/crowdsourcing-rdoc.shtml

Insel T, Cuthbert B, Garvey M, Heinssen R, Pine DS, Quinn K, Wang P. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. American Journal of Psychiatry. 2010; 167(7):748–751. [PubMed: 20595427]

Ioannidis JPA. Excess significance bias in the literature on brain volume abnormalities. Archives of General Psychiatry. 2011; 68(8):773–780. [PubMed: 21464342]

Ioannidis JPA. Research and theories on the etiology of mental diseases: doomed to failure? Psychological Inquiry. (this issue).

Kagan J. Amen. Psychological Inquiry. (this issue).

Klein DN, Hajcak G. Heterogeneity of depression: Clinical Considerations and Psychophysiological Measures. Psychological Inquiry. (this issue).

Levelt Committee, Noort Committee, & Drenth Committee. Stapel Investigation. 2015 from https://www.commissielevelt.nl/.

Lilienfeld SO, Pinto MA. Risky tests of etiological models in psychopathology research: The need for meta-methodology. Psychological Inquiry. (this issue).

Markon KE. Ontology, measurement, and other fundamental problems of scientific inference. Psychological Inquiry. (this issue).

Miller GA, Yee CM. Moving psychopathology forward. Psychological Inquiry. (this issue).

Nicholson JM, Ioannidis JPA. Research grants: Conform and be funded. [10.1038/492034a]. Nature. 2012; 492(7427):34–36. doi: http://www.nature.com/nature/journal/v492/n7427/abs/492034a.html#supplementary-information. [PubMed: 23222591]

NIH. NOT-OD-15-103: Enhancing Reproducibility through Rigor and Transparency. 2015. from http://grants.nih.gov/grants/guide/notice-files/NOT-OD-15-103.html

Nosek BA, Bar-Anan Y. Scientific Utopia: I. Opening cientific Communication. Psychological Inquiry. 2012; 23(3):217–243.

Polderman TJ, Benyamin B, de Leeuw CA, Sullivan PF, van Bochoven A, Visscher PM, Posthuma D. Meta-analysis of the heritability of human traits based on fifty years of twin studies. Nature Genetics. 2015

Regier DA. Potential DSM-5 and RDoC synergy for mental health research, treatment, and health policy advances. Psychological Inquiry. (this issue).

Szabo L. NIH director: Budget cuts put U.S. science at risk. 2014 from http://www.usatoday.com/story/news/nation/2014/04/23/nih-budget-cuts/8056113/.

The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. [10.1038/nature11632]. Nature. 2012; 491(7422):56–65. doi: http://www.nature.com/nature/journal/v491/n7422/abs/nature11632.html#supplementary-information. [PubMed: 23128226]

Widiger TA, Crego C, Oltmanns JR. The validation of a classification of psychopathology. Psychological Inquiry. (this issue).

Yong E. Psychologists strike a blow for reproducibility. Nature. 2013; 11:26.

Zachar P. Popper, Meehl, and progress: The evolving concept of risky test in the science of psychopathology. Psychological Inquiry. (this issue).

Table 1

Conducting Psychopathology Research: Problems and Solutions

What holds back progress?
• Overreliance on statistical modeling and technological innovation uninformed by causally-informative research design and plausible etiological theory
• Theory-derived confirmation bias that leads to analysis, presentation, and interpretation of results in a manner that favors the theory
• Publication bias and selective reporting of findings
• Lack of objective standard to determine what constitutes sufficient empirical support for valid interpretation of findings
• Overemphasis on novelty of findings to obtain funding or publication in high impact journal
• Emphasis on one approach to conceptualizing etiology/nosology to the exclusion of others
How can we move forward?
• Develop etiological theories and submit them to risky tests that narrow the number of interpretative possibilities
• Integrate results across methods and look for convergent findings, e.g., where a hypothesis generated in one domain can be tested in another
• Promote multiple theoretical perspectives and embrace contrary, critical, competitive peer review
• Value results that contradict theoretical expectations
• Value quality of theory, logic of design, and importance of mechanisms being tested to determine significance of findings
• Emphasize measurement quality and importance of discriminant as well as convergent validity
• Encourage large sample sizes, replicability, publication of null findings from adequately powered replication studies