Testing Scientific Claim At Own Risk: Reproducibility Against Novelty

Artem Akopyan

aakopyan@uwo.ca

Abstract: The article discusses the problem of validation by means of independent replication. Bem's studies of precognition are discussed in that context, as well as the recognized measure of implicit attitutes, the Affect Misattribution Procedure (AMP). Subsequently, a review of LeBel's replication of Payne, Hall, Cameron, and Bishara (2010) is presented. Finally, important replication-oriented initiatives are outlined.
[Artem Akopyan. **Testing Scientific Claim At Own Risk: Reproducibility Against Novelty**. *Rep Opinion* 2012; 4(12): 30-36]. (ISSN: 1553-9873). http://www.sciencepub.net/report. 6

Keywords: reproducibility, affect misattribution, questionable research

Theorizing and empirical research are often challenging, especially in the domain of social sciences where the omnipresent lack of truly objective perceptual evaluations is all the more apparent. Hypothesis testing is generally reduced to the aggregation of data scores and derivation of average descriptive values across the sample; this in turn makes research in psychology hypersensitive to blemishes in study design. For instance, Simmons, Nelson and Simonsohn (2011) demonstrated how selective reporting and optional stoppage can drastically inflate the probability of Type I error. Thus, the excessive use of researcher degrees of freedom might jeopardize the integrity and potential benefit of psychological science. Due to the reliance of inferential statistics on the likelihood of physical states, one of the verification (or falsification) procedures may involve the minimization of combined Type I error; in fact, the notion of reliability is well-known to psychologists, and every conclusion drawn from a set of data hinge upon the assumptions of validity and reliability of measurement instruments used, in other words that the test measured precisely the desired construct and that the ensuing conclusion would be found by any other researcher using the same or different sample from population. Despite the widespread awareness of the theoretical importance of reliability, in practice psychology researchers frequently display interpretation bias (LeBel & Peters, 2011; Wagenmakers, Wetzels, Borsboom, van der Maas, and Kievit, 2012). Since the primary goal of a psychologist is the detection of real world phenomena including covariation and cause-effect relations, the question of whether the same effect would be detected by an independent researcher is just as important as the surface appeal of that finding; in other words, independent replications allow one to convincingly argue the correctness of his or her hypothesis based on its consistent confirmation by colleagues. Without the concurrence in statistical results there may never be a reason for accepting the conclusion on faith; this is displayed perhaps most vividly in research topics involving processes that are implausible or have limited support in the scientific community.

Extrasensory perceptions (ESP) were investigated by Bem in a famous set of experiments (2011) where traditional statistical tests supported the hypothesis that participants were capable of anticipating the position of an erotic image prior to the latter's being displayed on screen. ESPs are generally attributed to the domain of parapsychology and are not widely acknowledged by scientists, and the statistical evidence is mixed at best. For instance, Wagenmakers et al. (2012) did not find sufficient evidence for precognition in an independent replication, allowing some of the skeptics to discard the phenomenon as unfounded. Dean Radin, however, notes that precognition does not contradict any of the known laws of nature accumulated thus far using the fundamental scientific principles; that is why outright rejection of precognition may be considered as short-sighted as its uncritical acceptance (Harvard University conference, 2011). For as long as humans experience unlikely sequences of events deriving from a gut feeling, prophetic dreams and the realization clairvoyant forecasts, the issue of psi will remain a subject of heated debate. There are less extreme examples of psychologists' indecision with regard to psychological phenomena, such as the internal mechanisms underlying measures of implicit attitudes. Implicit measures are at a disadvantage when pitted against their explicit counterparts in reliability estimates, due in part to the great complexity of the processes that elicit those attitudes and the likely less direct relation between the true implicit attitude and explicit judgment on a contrived experimental task. LeBel and Paunonen

(2011) demonstrated the tendency of low-reliability measures to have low reproducibility estimates. The logical implication of this relationship lies at the heart of scientific method: the power of an instrument to yield consistent results when applied to an external entity is an indication of that instrument's inherent stability. Among the most reputable implicit attitude measures are the IAT and AMP with the latter looking to replace the former as a reliable measure (Gawronski & Ye, 2012).

Payne et al (2005; 2010) proposed a process model of affect misattribution whereby a participant\s implicit attitudes are reflected by the misattribution of a prime's emotional valence onto evaluations of the target. Their model "…relies on the fact that people have difficulty disentangling their affective responses to two events occurring in close proximity in time and space. When this happens, people confuse the sources of their affective responses." (p.1398) Some of the experiments included explicit instruction for the participants to ignore the potential confound of the prime when a judgement about the target was being made. The stimuli presented were varied in pleasantness (e.g. puppy, snake) and participants' responses were fit to a multinomial model in which both correct responses and misattributions are represented by relative weights of true probabilities associated with a dichotomous left/right selection made at each stage of stimuli's presentation.

In line with AMP model, Oikawa, Aarts, and Oikawa's study (2011) showed how the misattribution was diminished when participants had to rate the primes for pleasantness prior to rating the target pictographs. They also suggested that AMP may be subject to pre-potent motor response patterns; Gawronski and Ye (2011), however, tested the latter confounding possibility rigorously and essentially refuted this suspicion. Nevertheless, there are other unresolved issues regarding the assumptions underlying the AMP, such as the frame of reference of pictographs. Payne, Cheng, Govorun, Stewart asked the participants to indicate whether each Chinese pictograph was "more or less pleasant than the average pictograph". Given that the aforementioned average pictograph (or a set of them) was never explicitly presented to participants, the request of comparing anything to the average could potentially be contaminated by participants' comparison with a previous pictograph as a frame of reference; participants' judgements therefore rested upon the agreement of their implicit perception of "the average pictograph" with that intended by the researchers, and most likely varied across the sample. Moreover, there was another inconsistency in the (self-contradictory) approach elected by the model's founders: while acknowledging the relation between large effect sizes and the affect's amplitude, the only distinction was between 0 and 1 for unpleasant and pleasant feelings, respectively; that is, no attempt was made to gauge the strength of the attitude.

LeBel (2012) conducted a replication of Payne, Hall. Cameron and Bishara (2010). The model proposed by Payne et al (2005; 2010) provided a good fit to the data (see Appendix). Across participants, the group of participants in the long presentation time condition showed equal amount of reduction in estimated mean scores compared to the short presentation group, in line with the notion of aggregate decrease in affect ratings when allowed to habituate to a stimulus. On the other hand, presentation time appeared to have weakened the effect of pictograph pleasantness on estimated marginal means across groups. If the AMP model were correct, one would perhaps expect the unwavering amplification of the intended pleasantness of the prime (increase in proportion of stimuli judged as pleasant for presumably pleasant pictographs, and a corresponding decrease in ratings of unpleasant pictographs). This finding is problematic as Payne, Cameron and Bishara recognize that primes were expected to be more influential with shorter presentation of stimuli. The results described above appear consistent with the idea of carryover from the perceptual evaluation assigned to the prime to that of the target. At longer presentation times, the participants were asked to evaluate the pictographs after a presumably greater reduction in the influence prime had on subsequent judgement. Under those circumstances, the estimated marginal mean were found to be much more level. Had the marginal mean estimate (MME) values displayed a trend of lower values for long presentation group for unpleasant and higher values for pleasant pictographs, Payne et al.'s model would have been entirely consistent with LeBel's (2012) replication data. Payne, Cheng, Govorun, and Stewart (2006) propose that simple affective reactions that have not been attributed to a particular source due to the lack of time elapsed since presentation, are the object of misattribution for a stimulus previously presented. Because the primes in Payne, Cameron, and Bishara's (2010) study were more familiar/readily recognizable than the unfamiliar Chinese pictographs, participants might have (hypothetically) been evaluating the primes and, upon being urged to report the affective response produced, rated the prime instead of the target pictograph. In that sense, the process of affect

misattribution resembles the predominant account of repetition blindness phenomenon in that the first stimulus presented, the prime, might yield stronger affective and/or semantic activation. Whether repetition blindness occurs is then presumably determined by the precise levels of activation elicited in each participant and temporal activation, among other possible factors (Morris, Still, and Caldwell-Harris, 2009); the AMP in its current state does not account for any of the listed variables. Moreover, the somewhat steeper negative slope across presentation time groups suggests that participants reacted to a prime as soon as it could be recognized. Negative stimuli preserved most of their impact as they informed the presence of danger in the participant's immediate environment; their positive counterparts did not have the same degree of evolutional underpinning and declined as participants processed them within the affective system (a positive stimulus under those experimental conditions is arguably more prone to ambivalence upon a revaluation: a puppy is not equally appealing to all participants after it had ruined a unique and/or expensive piece of furniture). Results obtained by LeBel thus correspond better to the model of explicit judgement of the prime in which participants evaluate the primes explicitly and simply report the affective response elicited by those primes; participants' ratings should in that case be considered fully processed emotions rather than simple affective reactions as intended by Payne, Cheng, Govorun and Stewart. Either the AMP model proved its adequacy in describing affect misattribution accurately and is a reliable measure as intended by Payne et al., or the primes were being evaluated (explicitly), and longer presentation time allowed participants to overcome the influence of simple affective reactions produced by the presumably more salient primes and yield a computed emotional response. The AMP model does not contradict either of the two mutually exclusive explanations and is ambiguous in that sense.

The author believes that neither hypothesis can be discarded without further independent replications. Bar-Anan and Nosek (2012), in fact, found evidence in favour of the explicit prime rating hypothesis stated above using self-report measures and "…without people who report that the attitude effect occurred, the psychometric qualities were very weak". Proponents of the AMP model may rightfully question the appropriateness of self-report methodology in the study of implicit measures, but Payne et al. (2005;2010) have used self-report methodology in their own studies of the AMP; more importantly, as has been previously noted, the model does not preclude a set of mutually conflicting explanations. Therein lay one of the most sublime shortcomings of contemporary psychological research: the lack of specificity in predicting the relationship between empirical support of a hypothesis and for those describing closely related processes. The Payne et al.'s theory depends on inaccuracy in perceiving the primary cause of a simple affect, as well as the ability of participants to evaluate sophisticated pictographs in a very short period of time; that is why one has to be cautious in embracing the theory until the research subfields of perception, general misattribution and emotion (in case participants were in fact evaluating the prime without attempting the reflection upon the unfamiliar pictographs). The AMP model is imprecise at least to the degree of supporting incompatible notions of the processes underlying affect misattribution. Thus, the model corresponds to a set of internal mechanisms, all of which are capable of producing the data found in the studies that tested the affect misattribution procedure. One theoretical ramification of the AMP model involves a further specification of expected outcomes with regard to related processes, such as attenuated attention, guessing and explicit rating of the prime. If researchers wishing to conduct exploratory research are required to make predictions about the data reflecting a host of related construct (such that all such constructs are inextricably tied to produce the effect studied by the researcher), there will be fewer papers submitted for publishing, and the ones that are offered to reviewers are rigorous in their conceptual carcass. This approach is impractical, however, as psychological journals are more interested in novel findings than confirmations of the soundness of known results (Nosek, Spies, and Motyl, 2012): more apparent becomes in psychological journals the phenomenon of publication bias (Francis, 2012; Giner-Sorolla, 2012).

The problem, however, is not as much with individual researchers using questionable research practices but rather in the incentive system currently in place. Articles that include more than one study have become more prevalent in psychological journals as the overall number of studies and theories offered to the scientific community has soared in the past few years, yielding a greater number of potential explanations of natural events, yet lacking definitiveness and empirical soundness. Strictly speaking, the stagnation in psychological science and the resulting praise of sensational findings have allowed a few researchers a rise to stardom in the academia by means of presenting hypotheses that are difficult to unequivocally refute. Such was the case of Diederik Stapel who had admitted to counterfeiting data for nearly thirty of his studies (Science, 2012).

The truth became known thanks to his colleagues who reported the fraud. Among the less blatant cases of the use of questionable research practices are the tendency of explaining nearly statistically significant results by random error and the multitude of internal processes that together comprise the cognitive systems of human beings; results that reach the .05 threshold are not scrutinized sufficiently as they are appealing and easy to process. If the social scientific community is motivated to conduct better research and report all the data and analyses openly without the pressure of publication hindering their own efforts, there will undoubtedly be fewer studies appearing in print but the overall value of published articles will increase dramatically.

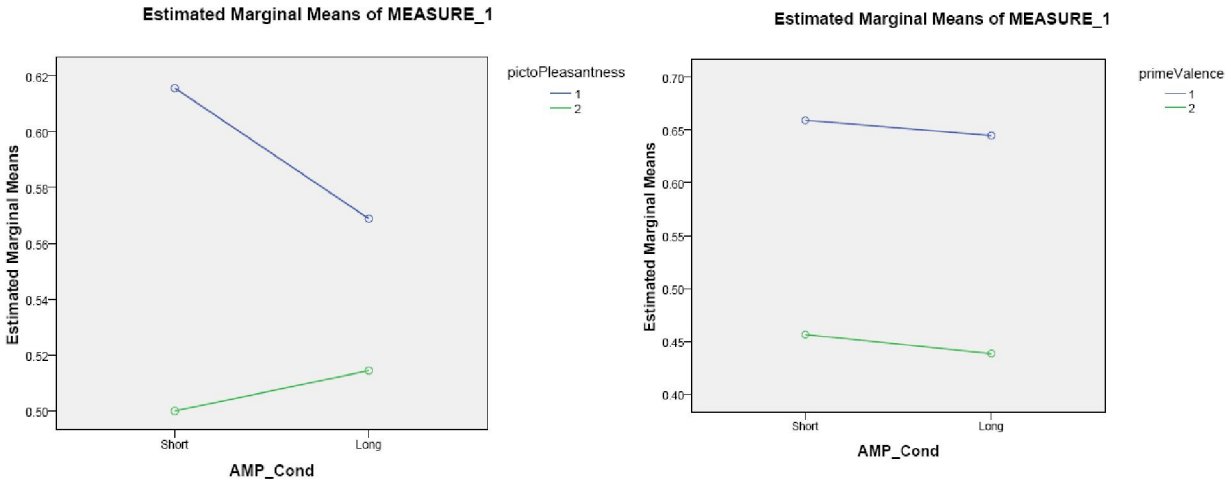
Most of the problems facing psychology at present result from the underestimation of the possibility of chance findings if they present the reader with a new perspective on the issue at hand; instead, more effort ought to be directed towards testing the already existent empirical results. The most commonly employed threshold is $\alpha = .05$, and results implying a Type I error probability of less than this value are readily submitted for publishing. Undergraduate disciples of psychological inquiry are taught null hypothesis significance testing (NHST) as the cornerstone of statistical analyses, following the cultivated yet arguably rudimentary tradition. Reliance upon NHST implies the dichotomy of support for the hypothesis or the failure of intended manipulation (Schimmack 2012). Masicampo and Lalande (2012), using probabilistic approach, compared the expected and observed frequency of published studies containing a p-value marginally lower than .05 and found the observed distribution highly unlikely provided psychologists publish all of their studies regardless of the computed p-value. Their study aptly demonstrated the condemnation of a principally outdated probabilistic convention with the use of probabilistic analyses. One of the problems with NHST, as pointed out by LeBel and Peters (2011), is the rigidity that is typical of modal research practices (MRP). Although some support is provided by a statistically significant result obtained in such a fashion, the hypothesis itself is far from unequivocal.

Admission requirements put in place for experimental "findings" are low enough to allow the acceptance of a theory on the basis of the original studies and a few conceptual replications. Wagenmakeers, Wetzels, Borsboom, van der Maas, and Kievit (2012) stress that researchers cannot be entrusted with observing the distinction between confirmatory and exploratory research thereby placing this responsibility upon the reviewers and authors' colleagues. One of the more progressive initiatives on that account has been the call for open communication between researchers and full disclosure of both the data collected and the statistical tests performed (Open Science Collaboration, 2012; Wicherts, Kievit, Bakker and Borsboom, 2012). Rather than impose the bottleneck of compliance with the demand for extraordinary results upon the researchers (which is unreasonable considering the bewildering complexity of personality pled by psychologists as an excuse for low correlation coefficients), psychologists should have a recognized alternative outlet for all of the studies being conducted, both those published in reputable journals and those rejected on account of a failure to reach statistical significance (Bakker, van Dijk, and Wicherts, 2012). Another well-known suggestion is the implementation of a radically new citation practices where psychology journals and online open access databases indicate with regard to an article the number of times each study in it has been literally replicated (this would encourage the popularity based upon replicable results rather than sensational conclusions derived from studies that nobody replicated exactly as intended by the original author; at the same time, researchers would be encouraged to concentrate on fewer studies but prepare them thoroughly without creating the illusion of an unequivocally sound theory). Reviewers themselves must be rated by psychologists and in this way motivated to be fair in assessments of submitted articles.

Interpretation bias (LeBel and Peters, 2011) is perhaps the most daunting challenge facing contemporary psychology as science is interested in the discovery of the ways in which nature operates as opposed to the ambiguity of initially promising results that are not supported by independent researchers. Bem argued at a conference (Psi and Psychology: The Recent Debate. 2011) that studies reporting statistically non-significant effect of precognition are not definitive in that a failed replication does not carry with it enough informational value to be per- or dissuaded of a phenomenon's existence in the real world. Perhaps even more telling is the terms "replication attempt" and "failed replication" as they are biased towards the initial experimenter's success and the replicator's failure due to random error or systematic differences introduced by the sample. Among the expository articles cited above, nearly all scrutinize articles that reported statistically significant results and express much less concern with hypotheses and theories that did not receive enough attention in the scientific circles due to the very same random error: even the

skeptics are driven mostly by the unfair disadvantage afforded by questionable research practices than the larger issue of self-correction of scientific enquiry. To this end, there is a need for consensus among social scientists with respect to basic assumptions shared by them; once all of the axiomatic concepts have been stated, research may be conducted and revised in accordance to the results' agreement with the axioms: while that would leave room for random error, investigative efforts would be directed towards the identification and quantification of constituent components of a theoretical construct, and the decision of whether a particular study is published would be founded in the contribution it makes to the inclusion/exclusion of a component from a potential enabling factor of a larger construct (for instance, if alcohol abuse is commonly found among individuals who are impulsive, prone to depression, and extraverted, the three subcomponents would be intensively studies by the scientists representing the research frontiers of personality psychology and psychopharmacology, to name a few). Schmidt (2009) points out the semantic distinction between exact replication and what is termed "close" (or "literal") replication: the former is impossible due to the use of different point in time for the experiment, as well as recruitment of different participants and/or different point in space; he points out that choosing to conduct a conceptual replication leads to the impossibility of falsifying a hypothesis by means of that replication as differences in procedures and materials may be consequential. Close replications minimize the imminent effects of systematic and unsystematic variation between samples in studies testing a given hypothesis or theory. Conceptual replications may therefore be seen as lending very limited support to a larger set of phenomena rather than testing the given hypothesis directly. As a result of over-reliance on conceptual replications, contemporary notions in psychological science were (are) being diluted to a vague conglomerate of effects and phenomena that are not questioned merely due to their outward plausibility; this is a major contributor to the stagnation of research in psychology (Carpenter, 2012).

What social sciences require in order for further progress in them to occur is a universal epistemological system that would inform subsequent investigation; the lack of infallible guidelines is the primary cause of the uncertainty surrounding contemporary psychology; in order for a system of this sort to function properly there must be a universal adherence to close replications as the most trustworthy tests of results published in social scientific journals. Unfortunately, close replications are for the most part rejected by reviewers, which effectually discourages falsification and with it scientific progress. Despite such discouraging outlooks among some of today's most revered psychologists, a growing number of psychologists recognize the importance of reporting findings that a colleague in a remote laboratory can replicate. The World Wide Web connects researchers throughout the world by Web-sites such as PsychFileDrawer, PLoS, and OpenScienceFramework. The Reproducibility Project, initiated by OSF is proving a success with inquisitive investigators pre-registering their methods, sampling, procedures, and reporting results with exactly the outlined statistical tests rather than ones that happened to produce an illusion of a true finding. Nonetheless, the vogue of reporting a set of underpowered studies has altogether not subsided and the notion of argumentative value of such publications must be disseminated. Furthermore, Frank and Saxe (2012) strongly recommended to University professors the practice of requiring close replications of their students, allowing the latter to internalize the scientific method while making a real contribution to psychology. Equally important is the explicit explanation to the students of the difference between exploratory and confirmatory research. Despite the hasty dismissal of close replications by reviewers and fellow scientists, psychologists must be urged to acknowledge that the potential detriment of random error in the so-called "failed replications" applies equally to the exploratory studies that draw extraordinary conclusions. There are enough exploratory findings in psychology as it is, and confirmatory research is on the agenda. Without the scrutiny of psychological findings by independent replications, the science of human behavior will remain a collection of enthralling stories that excite the reader but contradict one another and leave the inquisitive mind to explore the ways of science by trial and error, the probability of the latter likely being substantially higher than .05.



Appendix: SPSS output of data obtained by LeBel (2012) as part of the replication (kindly provided by Dr. Etienne LeBel).

References:

1. Bakker, M., van Dijk, A.& Wicherts, J.M. (2012). The rules of the game called psychologicalscience. Submitted to *Perspectives on Psychological Science.*
2. Bem, D.J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of personality and social psychology* 100, (3):407-425, https://www.lib.uwo.ca/cgi-bin/ezpauthn.cgi/docview/851236583?accountid=15115 (accessed December 17, 2012).
3. Blaison, C., Imhoff, R., Hühnel, I., Hess, U., & Banse, R. (2012, March 5). The Affect Misattribution Procedure: Hot or Not? *Emotion*. Advance online publication. doi:10.1037/a0026907.
4. Carpenter, S. (2012). Psychology's bold initiative. *Science* 336, (6076): 1558-1561, https://www.lib.uwo.ca/cgibin/ezpauthn.cgi/docview/1011858610?accountid=15115 (accessed December 17, 2012).
5. Francis, G. (2012). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic bulletin & review* 19, (2): 151-156, cgi-bin/ezpauthn.cgi/docview/1011861000?accountid=15115 (accessed December 17, 2012).
6. Gawronski B. & Ye, Y. (2012). What drives priming effects in the affect misattribution procedure? Underlying mechanisms and new applications. Submitted for publication.
7. Giner-Sorolla, R. (2012). Science or art? How esthetic standards grease the way through the publication bottleneck but undermine science. Submitted for *Perspectives on Psychological Science*.
8. John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science* 23, (5): 524-532, https://www.lib.uwo.ca/cgi-bin/ezpauthn.cgi/docview/1025879336?accountid=15115 (accessed December 17, 2012).
9. 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*
10. *Psychology Bulletin* 37, (4): 570-583, https://www.lib.uwo.ca/cgi-bin/ezpauthn.cgi/docview/865692448?accountid=15115 (accessed December 17, 2012).
11. Oikawa, O., Aarts, H., & Oikawa,H. (2011). There is a fire burning in my heart: The role of causal attribution in affect transfer. *Cognition and Emotion* 25, (1): 156-163, https://www.lib.uwo.ca/cgi-bin/ezpauthn.cgi/docview/852907790?accountid=15115(accessed December 17, 2012).
12. Masicampo, E. J., & Lalande, D.R. (2012): A peculiar prevalence of p values just below .05, The Quarterly Journal of Experimental Psychology, DOI:10.1080/17470218.2012.711335
13. Munafò, M. R. & and Flint, J. (2010). How reliable are scientific studies? *The British Journal of Psychiatry* 197, (4): 257-258, https://www.lib.uwo.ca/cgi-bin/ezpauthn.cgi/docview/900620505?accountid=15115 (accessed December 17, 2012).
14. Nosek, B.A. & Yoav Bar-Anan (2012). Scientific utopia: I. Opening scientific communication. *Psychological Inquiry* 23, (3): 217-243, https://www.lib.uwo.ca/cgi-

bin/ezpauthn.cgi/docview/1114699361?accountid=15115 (accessed December 17, 2012).

15. Nosek, B.A., Spies, J.R., & Motyl, M. (2012). Scientific utopia II: restructuring incentives and practices to promote truth over publishability. Submitted for Perspectives on Psychological Science.
16. Open Science Collaboration. (2012). An Open, Large-Scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science. Submitted for *Perspectives on Psychological Science*.
17. Payne, B. K., Cheng, C.M., Govorun, O., & Stewart, B.D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of personality and social psychology* 89, (3):277-293,https://www.lib.uwo.ca/cgi-bin/ezpauthn.cgi/docview/620950651?accountid=15115 (accessed December 17, 2012).
18. Payne, B. K., Hall, D.L., Cameron, C.D., & Bishara, A.J. (2010). A process model of affect misattribution. *Personality and Social Psychology Bulletin* 36, (10): 1397-1408, https://www.lib.uwo.ca/cgi-bin/ezpauthn.cgi/docview/815567678?accountid=15115 (accessed December 17, 2012).
19. Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological methods*, https://www.lib.uwo.ca/cgi-bin/ezpauthn.cgi/docview/1036890106?accountid=15115 (accessed December 17, 2012).
20. Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology* 13, (2): 90-100, https://www.lib.uwo.ca/cgi-bin/ezpauthn.cgi/docview/621988722?accountid=15115 (accessed December 17, 2012).
21. Simmons, J. P., Nelson, L.D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22, (11): 1359-1366, https://www.lib.uwo.ca/cgi-bin/ezpauthn.cgi/docview/912100859?accountid=15115 (accessed December 17, 2012).
22. Wagenmakers, E-J., Wetzels, R., Borsboom, D., van der Maas, H.J.L., & Kievit, R.A. (2012) - An agenda for purely confirmatory research. Submitted for *Perspectives on Psychological Science*.
23. Wagenmakers, E–J., Wetzels, R., Borsboom, D., & van der Maas, H.J.L. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of personality and social psychology* 100, (3): 426-432, https://www.lib.uwo.ca/cgi-bin/ezpauthn.cgi/docview/851235074?accountid=15115 (accessed December 17, 2012).
24. Wicherts, J.M. (2012). Letting the daylight in: reviewing the reviewers and other ways to maximize transparency in science. Frontiers in Computational Neuroscience, Vol. 6. doi: 10.3389/fncom.2012.00020
25. Yong, E. (2012). In wake of high-profile controversies, psychologists face up to problems with replication. *Nature*, Vol. 485

12/20/2012