Scott D. Churchill
Editor, The Humanistic Psychologist
Dept. of Psychology
University of Dallas
1845 E. Northgate Dr.
Irving, TX 75062

Dear Editor:

TOWARDS MORE EXPLICIT MEANS OF ASSESSING CONTROVERSIAL RESEARCH LITERATURES

In contrast with religious positions, humanistic viewpoints on life are not developed on the basis of sheer belief. A potential important source of knowledge consists of scientific data. Thus it is mandatory that a humanist has a reasonable perspective about the scientific literature that deals with questions about the fundamental aspects of human existence and behavior.

As Delgado-Romero and Howard point out in their article "Finding and Correcting Flawed Research Literatures" (*the Humanistic Psychologist*, 2005, **33 (4)**, 293–303) the scientific literature might contain results that seem to contain relevant answers to existential questions, but very probably are flawed. The scientific method is rather robust and in fields like physics one generally accepts that there flawed data (and theories!) abound. 'Time will tell' is the somewhat relaxed attitude there. Time in this case implies that (conceptual) replications of initially flawed results will not continue to repeat in the long run. In other fields with greater urgency like medicine, meta-analyses are often used to provide an evaluation, but Delgado-Romero and Howard feel correctly that selective reporting limits the use of meta-analyses for this goal. Publication bias tends to prolong the lifetime of flawed scientific data and by doing the replications oneself one can avoid this pitfall. In a remarkable article Ionnides for instance showed that over 50% of the significant findings in the psychological literature are not significant at all, but due instead to selective reporting (Ionnides, 2005).

To illustrate the laudable approach of 'doing your own set of replications' Delgado-Romero and Howard choose four fields of enquiry where controversial results were reportedly obtained. Not all of these fields claim to provide a new perspective on human existence. The first research field they applied their replication method to, of "Implementation Intentions," does not seem very relevant for humanistic thinking. The idea that music might influence cognitive performance (the Mozart effect) is the second of the potentially flawed results that they explore. That claim, if true, might provide a different view of the role of music in culture, and likewise it might have effects on human education, etc. But more to the point is the field of "the causal efficacy of remote intercessory prayer". Here the humanistic consequences are obvious. To evaluate the claim of efficacy they run three prayer studies of their own and find an effect size that differs considerably from the meta-analytic one, and is close to zero. Although it is unclear if their effect size differs significantly from the literature, they claim that this finding is sufficient to treat the literature on Intercessory Prayer with suspicion.

The final research field that they explore concerns the controversial topic of telepathy. Telepathy, if confirmed, would certainly have the potential to change our perspective on humankind. For instance, the prevailing scientific materialistic view with a focus on individualism might be replaced by a view allowing for some deep connections between humans.

Interestingly their treatment of the "potentially flawed data" in this case, in combination with the treatment of other fields, shows clearly that their evaluation method fails and is in need of a further, more quantitative, addition.

Delgado-Romero and Howard choose the ganzfeld telepathy paradigm for the replication efforts of telepathy because this paradigm provides the strongest cumulative evidence for an unexplained, anomalous effect, sometimes called "psi." In a ganzfeld study, participants are generally invited to the laboratory in pairs, where one is a "sender" and the other is a "receiver." Receivers are brought to a sound-proof room (one at a time), seated in a relaxing chair, the ears shielded by white noise on a headphone, the eyes shielded with goggles. They are instructed to relax and speak aloud a running account of their thoughts. Meanwhile, in another room, the sender concentrates on a randomly selected image or video-clip, trying to send it to their partner in the receiving room. At the end of the procedure, receivers are asked to pick which image or video-clip, out of four possibilities, was sent to them. The expected hit rate is thus 1 out of 4 (25%) if no psi is involved. Although three meta-analyses produced somewhat different results, (depending on inclusion criteria and weighting factors), the last and most comprehensive meta-analysis based on 79 Ganzfeld studies reported an average and highly statistically significant hit rate of 31% (Storm & Ertel, 2001). At this point it should be remarked that publication bias has been a source of concern in this research field, as in many others, but it has been evaluated and found to be negligible. Therefore we would expect that in the long run replication attempts should converge towards the meta-analytic estimate, and not towards chance. Delgado-Romero and Howard performed a total of eight ganzfeld studies, reported an average hit rate of 32%, which is statistically significant above chance and converges almost exactly towards the results found in the Storm and Ertel meta-analysis.

Following their own evaluation method the authors should have gracefully accepted that the research literature on the ganzfeld studies, at least until further evidence is produced, was apparently *not* flawed. Instead, they describe a "psychic theory," entailing that people are either fully psychic or not, and only pairs where both receivers and senders are psychic can make telepathy happen. Reasoning from this theory, the authors conducted yet another experiment selecting particularly good subjects (according to their theory) with an adapted procedure. Although the resulting hit rate was a surprising 13% (7 hits out of 52 trials), which is statistically significant *below* the expected 25% by chance, the authors state that based solely on this last experiment, "we do not believe that humans possess telepathic powers."

We would like to comment on the psychic theory proposed by Delgado-Romero and Howard, and on the application of their evaluation method.

A number of theories have been proposed to explain the anomalous results of the ganzfeld studies, but as far as we know, no one has been able to convincingly demonstrate that one theory is superior to the others. The theory proposed by Delgado-Romero and Howard belongs to a category of "mental radio" or signal-transmission models that were popular in the 1930s. Few parapsychological

researchers today believe that such theories are adequate for modelling the ganzfeld results. In fact, the ganzfeld effect may not involve telepathy at all, but rather precognition, where disclosing the correct target at the end of the session is somehow pre-sensed by the receiver. Also, parapsychological studies in general often show strong dependencies on the experimenter, where researchers following the exact same procedure can end up with results significantly coinciding with their own prior beliefs (Wiseman & Schlitz, 1997). This suggests that the presumed telepathy effect may be influenced, if not generated, by the people conducting the experiment.

The evaluation method for research literature proposed by Delgado-Romero and Howard has at least one further caveat. If one conducts a new series of studies to evaluate whether a research literature is flawed, one has to make choices concerning specific designs and procedures. A research literature combines studies that are similar, but often not in specific details. To provide a fair evaluation of the research literature, one must select designs and procedures that have consistently shown positive results. In the case of telepathy, the choice of the ganzfeld procedure may not have been the best choice for the topic of telepathy, but the choice was definitely fair in that the design and procedure of those studies show, on average, consistently above-chance hit rates. However, as soon as one departs from the accepted procedure, as the authors did in their last experiment, one is no longer evaluating the literature under scrutiny. In addition, if the test protocol has not been carefully described, it can be difficult to establish which details of the design and procedure are crucial, especially when it concerns controversial phenomena that are not well understood and where the important procedural details are even less well appreciated than in more conventional psychological tests.

Further, there is an inconsistency in the way Delgado-Romero and Howard evaluate the ganzfeld telepathy research and the way they evaluate the intercessory prayer and the other potentially literatures they consider. This inconsistency reveals an implicit prejudice, so we propose to improve their method by making such prejudices more explicit. In the case of the intercessory prayer literature Delgado-Romero and Howard stopped further evaluation after their first set of three replications showed a reason for suspicion. In the telepathy case they continued evaluating after eight replications continued to show results conforming closely to the meta-analytic finding. One might argue that this is justified on the basis of the following rule: “In case the results of an experiment confirm the meta-analytic estimates, one should continue to repeat with a new set of replications.” However, Delgado-Romero and Howard did not apply this rule in the case of the literature on “implementation intentions” because there they stopped immediately after finding that their own replication results were in line with the meta-analytic data.

There is a way to make sense of this inconsistent behavior. In their conclusions they recommend the use of Bayesian statistics. In that approach one formalizes one’s prior prejudices by setting a subjective *a priori* probability that the finding under study is true. For example, one could set *a priori* probabilities in favor of telepathy and the effects of intercessory prayer to be extremely low values, while one could set the probability for the reality of another controversial claim at higher values. By doing this, the reasons for why they evaluated the four research fields differently would be clearer.

We would like to recommend that Delgado-Romero and Howard's evaluation method be further extended by a quantitative treatment based upon Bayesian statistics. Specifically, it should include a stopping rule which specifies the number of replications one should conduct before halting further replications. This can be based upon a specified *a priori* probability that the effect is true, the effect size estimate of that effect, and most importantly, an explicit criterion for accepting the "truth" of the finding.

We think that the last experiment upon which Delgado-Romero and Howard decided to stop further evaluation is a valuable contribution to the research literature on psi, and it raises interesting questions: Why did that study generate results so significantly out of line with apparently similar studies? Can we identify discrepancies in design or procedures that account for this remarkable result? More fundamentally, can their finding be replicated, or does it turn out to be a fluke? Since the results of this one study are relatively extreme, one might even contend that this result can be explained by some kind of anti-psi, also known as "psi-missing" effect, in which one produces a result that is significantly below chance expectation.

In any case, their last study cannot form part of the evaluation method that the authors had proposed. First, it deviates too far from the protocol used in the studies the authors wished to evaluate. Second, conducting the final study with a different procedure implied that the first series of eight studies was somehow flawed, but the authors do not offer any suggestion of why those previous studies should be ignored. Third, the authors do not explain why the last study is suddenly decisive nor why they decided to stop after that one, especially because they proposed that one should conduct a series of studies to reach such conclusions (albeit with an unspecified stopping criteria). After all, they suspected the telepathy research literature to be flawed based on a file drawer effect, but basing a negative conclusion on a single negative outcome is as inadequate as selectively publishing only one positive finding out of series of findings, and then basing one's conclusion on that publication.

References

Ionnides, (2005). Why most Published Research Findings are False, *Plos* **2-8**, e124.

Storm, L. & Ertel, S. (2001). Does psi exist? Comments on Milton and Wiseman's (1999) meta-analysis of ganzfeld research. *Psychological Bulletin*, **127**, 424-433.

Wiseman, R., & Schlitz, M. (1997). Experimenter effects and the remote detection of staring. *Journal of Parapsychology*, **61**, 197-208.

Dick J. Bierman and Eva Lobach
Humanistic University
Van Asch van Wijckkade 28
Utrecht, The Netherlands