THE HUMANISTIC PSYCHOLOGIST, *33*(4), 293–303

Finding and Correcting Flawed Research Literatures

Edward A. Delgado-Romero
Department of Educational Psychology
University of Georgia

George S. Howard
Department of Psychology
University of Notre Dame

Humanistic psychology has always viewed scientific psychology with skepticism. Good reasons for this skepticism continuously appear. One is then left with the choice, "Is a scientific approach to humans inherently wrongheaded?" or "Is scientific psychology an imperfect but improving enterprise?" This article reviews another domain where research in scientific psychology proves misleading.

Suppose a psychologist was asked a question such as, "Is psychotherapy effective?" or "Is remote intercessory prayer effective?" or "Do humans possess psychic powers?" How might a psychologist reply? The most common strategy would be to conduct a meta-analysis over the relevant research literature and report the results. In all 3 cases (i.e., psychotherapy, efficacy of remote intercessory prayer, and telepathic powers) the answer would be a significant, positive effect size, suggesting that all 3 are real, efficacious phenomena. Unfortunately, in at least 2 of the 3 cases, the literature likely gives an incorrect answer to the question. How can one show that some literatures yield "incorrect" answers to research queries, whereas other literatures give "correct" answers? Finally, how should psychology's publication practices change to avoid flawed literatures?

For well over a century, a strain of thought in psychology has been skeptical of the scientific analysis of persons (Bakan, 1967; James, 1950). Around the middle of the last century, many of the voices of protest coalesced in the humanistic

Correspondence should be addressed to George Howard, Department of Psychology, University of Notre Dame, 218 Haggar Hall, Notre Dame, IN 46556. E-mail: Howard.2@nd.edu

psychology movement (Giorgi, 1970; Mair, 1989; Rogers, 1973; Wertz, 1992). A central tenet of humanistic psychology has been skepticism of a natural science approach to psychology, and the desire for a "human science" alternative (Giorgi, 1970, 1992). Although the development of a human science alternative has made some progress, humanistic psychology's larger impact has come as a critique of the natural science excesses of mainstream psychological research (Giorgi, 1970; Howard, 1982; Howard & Conway, 1986; Rogers, 1973). This article reviews yet another glaring problem with natural science research with humans and the need for (at least) a revision of psychology's current research practices.

In earlier programs of research, my colleagues and I found flaws in mainstream research methodologies regarding the value of retrospective pretests (e.g., Howard, Ralph, et al, 1979), use of behavior versus self-report measures (e.g., Howard, Maxwell, Wiener, Boynton, & Rooney, 1980), and the proportion of variance in human behavior due to free will (e.g., Howard & Conway, 1986). In each research program a "softer" or more humanistic alternative methodology was actually found that was more valid than a "harder" or more behavioral methodology that was favored in natural science, psychological research. The philosophy of science behind this empirical upgrading of accepted methodologies is laid out in Howard (1982).

A recent program of research (Howard, Hill, et al, 2005; Howard, Lau, et al, 2005; Lau, Howard, Maxwell, & Venter, 2005; Sweeny & Howard, 2005) suggested that there are problems with several research literatures in psychology. These problems are caused by the discipline's preference for significant, rather than nonsignificant, findings in deciding which articles will be published and which will not. This preference is often overtly stated, for example in the APA publication manual (2001) and in the editorial statements made by journal editors. It does not matter whether the decision to publish "good" (statistically significant results) or not to pursue or publish "bad" (nonsignificant findings) is made by a journal editor, a reviewer, or researchers themselves. It is problematic because any preference for "good" over "bad" findings leads to a biased—and sometimes severely misleading—research literature.

Imagine a baseball player who computes his batting average by only including days on which he got one or more hits (his good days) and eliminates days when he did not get any hits (his bad days). Obviously, his computed batting average (e.g., .600) would represent a gross distortion of his "real" (i.e., when all at bats are included) batting average (e.g., .300). This represents an analogy to the classic file drawer (Rosenthal, 1979) problem in psychological research. If one "reclaims" one half of the player's days from the file drawer, one could correct the misleading literature (e.g., .600) somewhat (e.g., batting average goes from .600 to .450). Although this corrected average is a more accurate estimate than the original batting average, it is still wrong (i.e., the "Truth" is .300, not .450).[1] Although all attempts to correct flawed meta-analyses yield improved estimates, unless one identifies all the studies in the "file drawer" (which is a virtual impossibility) the resulting estimate is "better," but still incorrect.

A METHOD FOR FINDING FLAWED LITERATURES—A NEW MOUSETRAP

Suppose one wanted to determine the extent to which the literature suggesting the effectiveness of psychotherapy is biased by the file drawer effect. The literature (Lambert & Bergin, 1994) suggests that the average treatment subject would be at the 80th percentile (instead of the 50th percentile) of a comparable control group at posttest. One way to test the validity of the psychotherapy effectiveness literature is to begin forming a new literature where there is no possibility of a file drawer effect. To do so, one must conduct several studies and accept all results (regardless of whether or not they reached statistical significance) that are obtained by methodologically adequate studies.

Imagine there are two different possibilities: the present literature is exactly correct (i.e., the average treatment subject percentile across new studies is .80, as Smith, Glass, and Miller, 1980 reported) or; the present literature is based solely on Type I errors (i.e., the average treatment percentile is .50, which would occur when the treatment and control groups means are identical). That is, imagine 100 studies were conducted and only 4 of them obtained significant findings. Further, the truth (if all 100 studies were published) is that treatment is not effective. However, if the literature consisted of only the four significant effects, which were the only studies published, then the literature would be based only on the four published Type I errors.

Now imagine that one conducted four studies (with a sample size of each study close to the average for the literature in question) and the treatment subjects' percentiles were .74, 62, .89, and .77. The average percentile of .76 looks very close to the literature's value of .80. Thus, one would tentatively conclude that the literature appears to be contaminated very little (or not at all) by the file drawer effect. This is because, if the null were true (i.e., 50th percentile), it is extremely unlikely that four independent studies would achieve results of .74, .62, .89, and .79. Still, our

[1] Meta-analysts have developed a number of adjustment methods in an effort to eliminate the effect of publication bias on estimating the true effect size (e.g., Duvall & Tweedie, 2000; Hedges & Vevea, 1996; Iyengar & Greenhouse, 1988). Although such methods can be helpful in reducing the effects of publication bias, current research suggests that even the most sophisticated of the methods developed so far cannot uniformly be relied on to eliminate bias (Kraemer, Gardner, Brooks, & Yesavage, 1998; Macaskill, Wallter, & Irwig, 2001).

conclusion must be tentative, because all the new studies would have been conducted by one research team. Thus, one would want to invite other researchers to replicate the findings in their labs, to insure that the data are representative.

Conversely, imagine the average percentile among treatment subjects in the four studies was .48 (e.g., .50, .46, .51, .45). Here one would tentatively conclude that the literature appears to be badly flawed, as these data are extremely unlikely if the real value was .80 (as the literature suggests). Again, further generalization replications would be required. Finally, one needs to conduct as many studies as are necessary to reasonably favor one outcome (i.e., .80, suggesting the existing literature is adequate) over the other (i.e., .50, the literature appears to be misleading).

BUT WILL IT CATCH MICE?

In an effort to test the proposed methodology for identifying flawed research literatures, we first considered the domain of implementation intentions (i.e., where people are forced to concretize, by writing them, their behavioral intentions; Gollwitzer, 1999). This literature was selected because we (Howard, Hill, et al., 2005) disagreed among ourselves as to the construct's likely efficacy (as opposed to the psychotherapy outcome literature, where we all thought that psychotherapy would be efficacious). The research literature on implementation intentions suggested an average effect size of $d = .54$ (a medium effect size, cf. Cohen, 1969) which was contrasted with a d of .00, which would imply the literature is solely the result of Type I errors.

Three studies testing the efficacy of implementation intentions relative to proper control groups found an average d score of .49. This finding was so close to our meta-analysis of the research literature ($d = .54$) that it seems the existent literature is relatively free of debilitating file drawer effects (as $d = .00$ is very unlikely). Although it was reassuring to find that a literature stood up well to our scrutiny, we desired to see if the methodology could identify a literature that would not stand up to scrutiny (i.e., a flawed literature).

The next construct that the team members disagreed on was the emerging literature that suggests the causal efficacy of remote intercessory prayer (Harris et al., 1999). The meta-analysis revealed a small but significant effect size of $d = .21$. Our three new studies averaged an effect size of $d = .02$, suggesting that the existing literature might be based on Type I errors, because $d = .02$ is clearly closer to $d = .00$ than it is to $d = .21$. Of course, replications of our findings by other researchers would be required before stating a definitive conclusion. However, our technique seems to be able to pick up problematic research literatures. Thus, it seems likely that the many psychologists who tout the proven efficacy of remote intercessory prayer are basing their claims on a flawed, misleading literature.

The remaining constructs to be tested were selected because we suspected the existing literatures might be misleading. The *Mozart effect* suggests that participants score higher on certain IQ subscales immediately after listening to a Mozart sonata rather than when participants are exposed to a variety of attention placebos. The most recent meta-analysis (Hetland, 2000) suggested an average effect size of *d* = .46 for the Mozart effect. Three studies were conducted and yielded an average effect size of *d* = –.14 (Sweeny, 2005). Again a *d* = –.14 is much closer to *d* = .00 than it is to the literature's value of *d* = .46. Thus, our data suggest that the existing literature on the Mozart effect is also problematic. Again, further generalization studies are required before stating strong conclusions.

To this point, only three studies have been required to render tentative decisions on the adequacy of various research literatures. Our last research literature (Howard, Lau, et al., 2005; Lau et al., 2005) required considerably more than three studies to untangle. The question was, "Do humans possess telepathic powers?" Although most psychologists believe humans are not psychic, the research literature argues strongly that humans are psychic (see Bem & Honorton, 1994, and Storm & Ertel, 2001, for positive meta-analyses). However, even after more than 80 years of research, the issue remains unresolved (see Milton & Wiseman, 1999, for a negative meta-analysis). First, we need to explain a bit of why the research picture is currently so unclear.

The strongest methodological procedure for testing humans' telepathic power is the ganzfeld procedure, although the literature includes many other approaches for testing humans' psychic powers. In the ganzfeld procedure, participants try to choose a target picture out of an array of four pictures as the dependant variable. Thus, participants must beat a precise chance level (25%) of correct choices to demonstrate psychic ability. For clarity's sake, all results will be expressed as percentage correct scores rather than *d* scores. The two positive meta-analyses found percent correct responses of .32 (Bem & Honorton, 1994) and .31 (Storm & Ertel, 2001). These findings suggest that a few people (but a statistically significant number) possess telepathic powers. But, because these psychics are mixed in with the 25% who are not psychic, but who correctly identify the target picture by chance, identification of the true psychics (if such people exist) is not possible.

The negative meta-analysis (Milton & Wiseman, 1999) found a hit rate of 27% where results from only the methodologically stronger ganzfeld procedure were included.[2] One should note that the results of the negative meta-analysis and the positive meta-analyses were quite close to one another (from 27% to 32%).

[2] Milton and Wiseman (1999) employed a rather puzzling form of meta-analysis. They reviewed 30 studies and weighted each study equally (i.e., unweighted procedure), rather than the widely accepted procedure of weighting each study by the study's sample size (i.e., weighted procedure). Thus, one study that ran only 4 pairs of subjects received the same weighting in the overall Effect Size (ES) as did another that tested 100 pairs. Using the more common weighted procedure, the mean ES more than doubled from .013 to .028.

In our first study (Lau et al., 2005), we found that a significant (45%) group of participants chose the correct target. In the second study, 40% chose the correct picture, and 20% correct choices were made in our third study. Although the percent correct hits across the three studies (i.e., 35%) was greater than even the positive meta-analyses, this overall percentage was not significantly greater than chance. These conflicting results led us to conduct five more studies, obtaining correct percentages of 30%, 30%, 20%, 35%, and 40%. After eight studies, we had an overall hit rate of 32% (which agrees with the positive meta-analyses) and, in fact, our hit rate was also statistically significant, $\chi^2(1) = 4.03$, $p < .05$. Further, when our data are added to the Milton and Wiseman (1999) meta-analysis over ganzfeld studies, the overall percent correct responses goes from 26% to 27% and this value now is very close to significant. So, for the moment, even the evidence against humans possessing psychic powers is precariously close to demonstrating humans do have psychic powers. The lower boundary of the confidence interval is now 24.7%, which is extremely close to not including the 25% value.

AFTER FURTHER REVIEW …

In the National Football League, the referees call every play and their decision on the field stands—unless it is challenged by one of the team's coaches. The head referee then looks at the challenged play on videotape (which provides a very different perspective on the play), and the referee uses the videotape to determine if there is sufficient evidence on the videotape to warrant overturning the call made on the field. Because no humans are perfect, it is wise to periodically overturn decisions, if the evidence from a somewhat different perspective warrants the change. Referees announce their decision, to either uphold their call on the field or to reverse themselves, with the phrase, "After further review... ."

To this point, our research team's data suggested that the literatures on remote intercessory prayer and the Mozart effect were problematic. We are comfortable with those conclusions, and hope generalization studies will begin immediately. Our data also suggested that the implementation intentions literature is solid, and we see no reason to further test that construct. Finally, although our own data (and the data of many others) suggest that some humans possess psychic powers, the team is still very uncomfortable with that conclusion. Thus, we initiated a further review of the ganzfeld procedure and the data obtained when using it.

In the ganzfeld procedure, participants are run in pairs. According to psychic theory, if one member of the pair is psychic (P) but the other is not (n) there will be no transmission of information. Only PP pairs can successfully send and receive messages. What exactly does this imply?

If one were to obtain only a chance level of correct responses, one would have five chance hits in every 20-pair study. What number of correct hits would be expected if every participant was a P? Here, all pairings would be PP, so the expected number of correct choices would be 20 (assuming PP pairs always correctly send and receive messages).

Now consider an intermediate situation. Imagine one in four people are psychic, which is an interesting possibility, and that Ps always successfully send and receive messages. What would be the expected value for a 20-pair study? It would be less than 20, but more than 5, obviously. Exactly what would one think the expected value would be?

It turns out the expected value is a paltry 5.95, and all the data in the telepathy research obtains values of this magnitude. Why does the expected value fall so precipitously from 20 (when all are Ps) to only 5.95 (when 25% are Ps)? The probability of the sender being psychic is 1/4 and the probability of the receiver being a P is also 1/4. It follows that there is only 1 PP pair in every 16 pairs! Thus, in a 20-pair study, one would expect (on average) five correct chance hits plus 1.25 PP hits, minus the likelihood that the chance hits might involve a PP pair (or .30). Why is this dramatic drop in the expected value for the ganzfeld procedure so important? It is because it makes a very small deviation from chance (5.95 – 5. = .95) all that one can reasonably hope to obtain, when an interesting phenomenon (25% of participants are psychic) is posited. The interpretation of studies with low statistical power is problematic. We now have unearthed a methodological parallel—the ganzfeld procedure is (methodologically) far less powerful than any of us had imagined. Finally, this reduced methodological power is caused by too low a ratio (1/4) of Ps to total participants. Is there any way of enriching this ratio?

Shortly after our sixth telepathy study, we hit upon a method for increasing the proportion of Ps in our sample. Suppose we considered only the pairs that chose the correct picture in our last two studies as participants in a different kind of study. What would be the ratio of P to total participants in that group? If the Ps were 25% in the original group, there would be 43% Ps in the group that answered correctly. This means that more than one in four pairs are PP in this enhanced sample. This represents a four-fold improvement on the 1/16 likelihood of PP pairs in the original studies.

We also realized that our confidence in our findings would be greatly enhanced if pairs went through the ganzfeld procedure four times, instead of just once. We paid pairs $100 if they would participate in the second part of our study. Out of the final 40 pairs, 15 pairs chose the correct picture. Of these 15 pairs, 13 pairs agreed to participate in a second (four-trial) study.

Because each pair had a one-in-four correct by chance probability in the four-trial study, exactly one correct response was expected by chance. Pairs that got zero or one (chance level) pictures correct would (likely) not be PP pairs. Pairs that got two correct might be a PP pair, but also might be a pair that was slightly luckier than chance. Any pairs getting three or four correct pictures (out of four) would likely be PP pairs. If the null was true, the number of pairs who got two or more correct answers should have equaled the number of pairs who got zero correct (because one correct choice is the chance level).

Of our 13 pairs, there were no fours or threes; 1 pair got two correct; 5 pairs had one correct picture; and 7 pairs had no correct responses. These are enormously disappointing data for individuals who believe humans possess psychic powers—especially because the sample had undergone a selection procedure to increase the percentage of Ps in the sample. Due to this last data set, we do not believe that humans possess telepathic powers. Further, the approximately 32% correct figure obtained in an enormous number of psi studies remains perplexing. Perhaps this 7% phenomenon is comparable to Meehl's (1978) "crud factor," which suggests that everything is correlated with everything else to a small degree. Meehl cited this as evidence that a null hypothesis is never literally true. Or, perhaps it simply reflects our enduring preference for significant results over studies that obtain nonsignificant findings.

CONCLUSIONS

Sometimes it is necessary for a science to first take a step backward, then take two steps forward. Psychology's research literatures now seem to be more problematic than most had imagined. This difficulty is caused by the discipline's preference for significant findings over nonsignificant effects when deciding which studies will be published and which will be discarded. This situation is clearly a step backward for the discipline.

The procedure described herein represents a promising technique for sorting adequate research literatures (i.e., implementation intentions) from the more problematic research literatures (i.e., remote intercessory prayer, the Mozart effect). And in some instances (i.e., psychic powers) the answers still are a bit unclear. This new methodology appears to be a promising method for sorting the grain from the chaff. Are there any ways to construct research literatures that will not be problematic?

There are two very different solutions to this problem. The first solution has the benefit of allowing us to keep our classical null hypothesis testing statistics. Researchers would submit only the introduction and methods sections of their articles. Reviews of these sections determine whether the results (whatever they might be) will be published. This procedure would give an equal likelihood of being published to good (significant) and bad (nonsignificant) results.

The second solution is actually better, but it requires greater disruption of psychology's publishing process. We recommend that Bayesian statistics replace the current classical (Fischerian or frequentist) statistical techniques. Any scientist is most interested in the question "What is the likelihood that my theory (e.g., that humans have psychic powers, that remote intercessory prayer produces real effects, etc.) is true, given my data?" That would be the probability of my theory given the data (T/D). Is that what our statistical techniques now compute? Absolutely not! They calculate the probability of D/not T (this data set given the null).

Given that our statistical techniques calculate something other than what scientists desire (T/D), and because they are constructed to make binary decisions about possibly true/possibly false null hypotheses, psychologists should seriously consider Bayesian statistics. First they compute the probability of T/D, which scientists have always desired. Second, they are unconcerned with accepting or rejecting null hypotheses; and so acceptance or rejection of studies would be based solely upon methodological considerations. Thus, the issue of problematic literatures (like those noted herein) is moot. Bayesian statistics are designed to fix effect sizes (e.g., humans select 32% correct pictures when 25% are expected by chance alone; implementation intentions produce an effect size of $d = .54$, etc.) and completely avoid the problems of accepting or rejecting studies based on the research's outcomes. One can get a good feel for how Bayesian statistics compare with classical statistical techniques from Howard, Maxwell, and Fleming (2000), where multiple data sets are analyzed with Bayesian and classical statistical techniques. Although it is always frustrating to take a step backward, sometimes it is a necessary task for any science. Our hope is that this small step backward will enable psychology to take its next steps forward.

Finally, these conclusions assume that a natural science approach is flawed but improvable (Howard, 1982, 1986). Some in the human science tradition would see the ambition of this series of studies as a "fool's errand," because such thinkers believe the natural science approach is inherently flawed. Be that as it may, this line of studies has helped mainstream psychology to recognize a prevalent flaw in the way it uses Fisherian statistics. It holds hope that psychology might someday become a better science of humans.

REFERENCES

American Psychological Association. (2001). *Publication Manual of the American Psychological Association* (5th ed.). Washington, DC: APA Books.

Bakan, D. (1967). *On method: Toward a reconstruction of psychological method*. San Francisco: Jossey-Bass.

Bem, D. J., & Honorton, C. (1994). Does psi exist? Replicable evidence for a anomalous process of information transfer. *Psychological Bulletin, 115,* 4–18.

Byrd, R. C. (1988). Positive therapeutic effects of intercessory prayer in a coronary care unit population. *Alternative Therapies in Health and Medicine, 3,* 87–90.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Duvall, S., & Tweedie, R. (2000). A non-parametric "trim and fill" method of assessing publication bias in meta-analysis. *Journal of the American Statistical Association, 95,* 89–98.

Giorgi, A. (1970). *Psychology as a human science*. New York: Harper & Row.

Giorgi, A. (1992). The idea of human science. *The Humanistic Psychologist, 20,* 202–217.

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist, 54,* 493–503.

Harris, W. S., Gowda, M., Kolb, J. W., Strychacz, C. P., Vacek, J. L., Jones, P. G., et al. (1999). A randomized, controlled trial of the effects of remote, intercessory prayer on outcomes in patients admitted to the coronary care unit. *Archives of Internal Medicine, 159,* 2273–2278.

Hedges, L. V., & Vevea, J. L. (1996). Estimating effect size under publication bias; Small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics, 21,* 299–332.

Hetland, L. (2000). Listening to music enhances special-temporal reasoning: Evidence for the Mozart effect. *Journal of Aesthetic Education, 34,* 105–148.

Howard, G. S. (1982). Improving methodology via research on research methods. *Journal of Counseling Psychology, 29,* 318–326.

Howard, G. S. (1986). *Dare we develop a human science?* Notre Dame, IN: Academic Publications.

Howard, G. S., & Conway, C. G. (1986). Can there be an empirical science of volitional action? *American Psychologist, 41,* 1241–1251.

Howard, G. S., Hill, T. L., Maxwell, S. E., Baptista, T. M., Farias, M. H., Coelho, C., et al. (2005). *What is wrong with research literatures? And how to make them right.* New Ideas in Psychology (under review).

Howard, G. S., Lau, M., Johnson, I., Johnson, M., Morrow, V., Looman, M., et al. (2005). Should two wrongs make a right? *Review of General Psychology.*

Howard, G. S., Maxwell, S. E., & Fleming, K. J. (2000). The proof of the pudding: An illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. *Psychological Methods, 5,* 315–332.

Howard, G. S., Maxwell, S. E., Wiener, R. L., Boynton, K. S. & Rooney, W. M. (1980). Is a behavioral measure the best estimate of behavioral parameters? Perhaps not. *Applied Psychological Measurement, 4,* 63–69.

Howard, G. S., Ralph, K. M., Gulanick, N. A., Maxwell, S. E., Nance, D. W., & Gerber, S. L. (1979). Internal invalidity in pretest–posttest self-report evaluations and a re-evaluation of retrospective pretests. *Applied Psychological Measurement, 3,* 1–23.

Iyengar, S., & Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science, 3,* 109–135.

James, W. (1950). *Principles of psychology*, New York: Dover.

Kraemer, H. C., Gardner, C., Brooks, J. O., & Yesavage, J. A. (1998). Advantages of excluding underpowered studies in meta-analysis: Inclusionist versus exclusionist viewpoints. *Psychological Methods, 3,* 23–31.

Lambert, M. J., & Bergin, A. E. (1994). The effectiveness of psychotherapy. In S. Garfield & A. E. Bergin (Eds.), *The handbook of psychotherapy and behavior change* (pp. 143–189). New York: Wiley.

Lau, M., Howard, G. S., Maxwell, S. E., & Venter, A. (2005). A method for settling research controversies: Are humans telepathic. *Psychological Bulletin* (under review).

Macaskill, P., Walter, S., & Irwig, L. (2001). A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine, 20,* 641–654.

Mair, J. M. M. (1989). *Between psychology and psychotherapy: The poetics of experience*. London: Routledge.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.

Milton, J., & Wiseman, R. (1999). Does psi exist? Lack of replication of an anomalous process of information transfer. *Psychological Bulletin, 125,* 387–391.

Rogers, C. R. (1973). Some new challenges. *American Psychologist, 28,* 379–387.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86,* 638–641.

Smith, M. L., Glass, G. V., & Miller, T. I. (1980). *The benefits of psychotherapy.* Baltimore, MD: Johns Hopkins University Press.

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.

Sweeny, R. (2005). A test of the validity of the "Mozart effect." Unpublished Doctoral dissertation, University of Notre Dame, Notre Dame, IN.

Wertz, F. J. (1992). The humanistic movement in psychology: History, celebration and prospectus. *The Humanistic Psychologist, 20*(2&3), 124–476.

AUTHOR NOTE

Edward A. Delgado-Romero is an Associate Professor of Counseling Psychology at the University of Georgia. He received his PhD in counseling psychology from the University of Notre Dame.

George S. Howard is the Morohan Director of the Arts and Letters College Seminar and a Professor of Psychology at the University of Notre Dame. He received his PhD in counseling psychology from Southern Illinois University, Carbondale.

Authors of many books, their most recent collaboration is entitled, *When Things Begin to Go Bad* published by Hamilton Books, Lantham, MD, 2004.