Psychological Bulletin
1994, Vol. 115, No. 1, 25–27

Copyright 1994 by the American Psychological Association, Inc.
0033-2909/94/$3.00

Response to Hyman

Daryl J. Bem

R. Hyman (1994) raises two major points about D. J. Bem and C. Honorton's (1994) article on the psi ganzfeld experiments. First, he challenges the claim that the results of the autoganzfeld experiments are consistent with the earlier database. Second, he expresses concerns about the adequacy of the randomization procedures. In response to the first point, I argue that our claims about the consistency of the autoganzfeld results with the earlier database are quite modest and challenge his counterclaim that the results are inconsistent with it. In response to his methodological point, I present new analyses that should allay apprehensions about the adequacy of the randomization procedures.

I am pleased that Ray Hyman, one of parapsychology's most knowledgeable and skeptical critics, concurs with Charles Honorton and me on so many aspects of the autoganzfeld experiments: the soundness of their methodology, the clear rejection of the null hypothesis, and, of course, the need for further replication. I hope this brief response will further augment our areas of agreement.

Hyman raises two major points about our article. First, he challenges our claim that the results of the autoganzfeld studies are consistent with those in the earlier database. Second, he expresses concerns about the "incomplete justification of the adequacy of the randomization procedures" and speculates that inadequate randomization may have interacted with subject or experimenter response biases to produce artifactual results.

Consistency With the Earlier Database

The earlier ganzfeld database comprised studies whose methods and results were quite heterogeneous. Consequently, one cannot justify any strong claims that some subsequent finding is either consistent or inconsistent with that database. For this reason, Honorton and I were careful not to make such claims. With regard to the major finding, we simply observed that earlier studies had achieved an overall hit rate of about 33% (25% would be expected by chance) and noted that the autoganzfeld experiments achieved approximately the same effect size. End of claim.

In general, the earlier database served primarily to suggest the kinds of variables that needed to be examined more systematically or more rigorously in the new studies. For example, previous ganzfeld studies that had used multi-image View Master slide reels as target stimuli obtained significantly higher hit rates than did studies that had used single-image photographs. This finding prompted Honorton and his colleagues to include both video film clips and single-image photographs in the autoganzfeld experiments to determine whether the former were superior. They were. Our only claim about methodological comparability was the modest observation that "by adding motion and sound, the video clips might be thought of as high-tech versions of the View Master reels."

But Hyman argues at length that video clips are not *really* like View Master reels. Surely this is a matter of interpretation, but does it really matter? Usually in psychology, successful conceptual replications inspire more confidence about the reality of the underlying phenomenon than do exact replications. I believe that to be the case here.

An example of a variable selected from the earlier database for more rigorous reexamination was sender–receiver pairing. Previous ganzfeld studies that permitted receivers to bring in friends to serve as senders obtained significantly higher hit rates than did studies that used only laboratory-assigned senders. But as we emphasized in our article, "there is no record of how many participants in the former studies actually brought in friends," and hence these studies do not provide a clean test of the sender–receiver variable. Moreover, the two kinds of studies differed on many other variables as well.

In the autoganzfeld studies, all participants were free to bring in friends, and it was found that sender–receiver pairs who were friends did, in fact, achieve higher hit rates than did sender–receiver pairs who were not friends (35% vs. 29%). But the reliability of this finding is equivocal. In the archival publication of the autoganzfeld studies, Honorton et al. (1990) presented this finding as a marginally significant point-biserial correlation of .36 (p = .06). In our article, however, we chose to apply Fisher's exact test to the hit rates themselves. Because this yielded a nonsignificant p value, we thought it prudent simply to conclude that "sender–receiver pairing was not a significant correlate of psi performance in the autoganzfeld studies."

But to Hyman, "this failure to get significance is a noteworthy inconsistency." (In part, he makes it appear more inconsistent than it is by erroneously stating that the earlier database yielded a significant difference in performance between friend pairs and nonfriend pairs. As noted earlier, this is an indirect inference at best.)

I am grateful to Richard Broughton of the Institute for Parapsychology in Durham, North Carolina, for going through the original autoganzfeld computer files with me to unearth the data necessary for the additional analyses presented in this response.

Correspondence concerning this article should be addressed to Daryl J. Bem, Department of Psychology, Uris Hall, Cornell University, Ithaca, New York 14853. Electronic mail may be sent to d.bem@cornell.edu.

I submit that Hyman is using a double standard here. If the successful replication of the relation between target type and psi performance is not analogous to the earlier finding with the View Master reels, then why is this near miss with a methodologically cleaner assessment of the sender–receiver variable a "noteworthy inconsistency"?

Hyman cannot have it both ways. If the heterogeneity of the original database and the methodological dissimilarities between its variables and those in the autoganzfeld studies preclude strong claims of consistency, then these same factors preclude strong claims of inconsistency.

Randomization

As we noted in our article, the issue of target randomization is critical in many psi experiments because systematic patterns in inadequately randomized target sequences might be detected by subjects during a session or might match their preexisting response biases. In a ganzfeld study, however, randomization is less problematic because only one target is selected during the session and most subjects serve in only one session. The primary concern is simply that all the stimuli within each judging set be sampled uniformly over the course of the study. Similar considerations govern the second randomization, which takes place after the ganzfeld period and determines the sequence in which the target and decoys are presented to the receiver for judging.

In the 10 basic autoganzfeld experiments, 160 film clips were sampled for a total of 329 sessions; accordingly, a particular clip would be expected to appear as the target in only about 2 sessions. This low expected frequency means that it is not possible to statistically assess the randomness of the actual distribution observed. Accordingly, Honorton et al. (1990) ran several large-scale control series to test the output of the random number generator. These control series confirmed that it was providing a uniform distribution of values through the full target range. Statistical tests that could legitimately be performed on the actual frequencies observed confirmed that targets were, on average, selected uniformly from among the four film clips within each judging set and that the four possible judging sequences were uniformly distributed across the sessions.

Nevertheless, Hyman remains legitimately concerned about the adequacy of the randomizations and their potential interactions with possible receiver or experimenter response biases. Two kinds of response bias are involved: differential preferences for video clips on the basis of their content and differential preferences for clips on the basis of their position in the judging sequence.

Content-Related Response Bias

Because the adequacy of target randomization cannot be statistically assessed owing to the low expected frequencies, the possibility remains open that an unequal distribution of targets could interact with receivers' content preferences to produce artifactually high hit rates. As we reported in our article, Honorton and I encountered this problem in an autoganzfeld study that used a single judging set for all sessions (Study 302), a problem we dealt with in two ways. To respond to Hyman's concerns, *I have now performed the same* two analyses on the remainder of the database. Both treat the four-clip judging set as the unit of analysis, and neither requires the assumption that the null baseline is fixed at 25% or at any other particular value.

In the first analysis, the actual target frequencies observed are used in conjunction with receivers' actual judgments to derive a new, empirical baseline for each judging set. In particular, I multiplied the proportion of times each clip in a set was the target by the proportion of times that a receiver rated it as the target. This product represents the probability that a receiver would score a hit if there were no psi effect. The sum of these products across the four clips in the set thus constitutes the empirical null baseline for that set. Next, I computed Cohen's measure of effect size (h) on the difference between the overall hit rate observed within that set and this empirical baseline. For purposes of comparison, I then reconverted Cohen's h back to its equivalent hit rate for a uniformly distributed judging set in which the null baseline would, in fact, be 25%.

Across the 40 sets, the mean unadjusted hit rate was 31.5%, significantly higher than 25%, one-sample $t(39) = 2.44, p = .01$, one-tailed. The new, bias-adjusted hit rate was virtually identical (30.7%), $t(39) = 2.37, p = .01, t_{\text{diff}}(39) = 0.85, p = .40$, indicating that unequal target frequencies were not significantly inflating the hit rate.

The second analysis treats each film clip as its own control by comparing the proportion of times it was rated as the target when it actually was the target and the proportion of times it was rated as the target when it was one of the decoys. This procedure automatically cancels out any content-related target preferences that receivers (or experimenters) might have. First, I calculated these two proportions for each clip and then averaged them across the four clips within each judging set. The results show that across the 40 judging sets, clips were rated as targets significantly more frequently when they were targets than when they were decoys (29% and 22%, respectively), paired $t(39) = 2.03, p = .025$, one-tailed. Both of these analyses indicate that the observed psi effect cannot be attributed to the conjunction of unequal target distributions and content-related response biases.

Sequence-Related Response Bias

Hyman is also concerned about the randomization of the judging sequence

> because we can expect strong systematic biases during the judging procedure. The fact that the items to be judged have to be presented sequentially, when combined with what we know about subjective validation . . . would lead us to expect a strong tendency to select the first or second items during the judging series.

Hyman's hypothesis is correct: As shown in Table 1, receivers do display a position bias in their judgments $\chi^2(3, N = 354) = 8.64, p < .05$, tending to identify as targets clips appearing either first or last in the judging sequence. Moreover, the actual distribution of targets across the judging positions also departs significantly from a uniform distribution, $\chi^2(3, N = 354) = 7.83, p < .05$, with targets appearing most frequently in the third position.

To determine whether the conjunction of these two unequal distributions might contribute artifactually to the hit rate, one can again combine the observed frequencies to derive an empirical null baseline. As shown in Table 1, each proportion in the second column can be multiplied by the corresponding proportion in the third column to yield the hit rate expected if there were no psi effect. As shown, the expected hit rate across all four judging positions is 24.7%.

The pertinent fact here is that this is lower than the 25% that would have been obtained if the target positions had been uniformly distributed across the sessions. In other words, the conjunction of receivers' position biases with the imperfect randomization of target positions works *against* successful psi performance in these data. Again, inadequate randomization has not contributed artifactually to the hit rates.

Table 1
Proportion of Sessions in Which Each Clip Was Selected as the Target and Proportion in Which It Appeared as the Target

Position in judging sequence	Selected as target	Appeared as target	Expected hit rate (%)
1	.30	.25	7.5
2	.20	.24	4.9
3	.22	.31	6.7
4	.28	.20	5.6
Total	1.00	1.00	24.7

Note. $N = 354$ sessions.

Alternative Randomizing Strategies?

Hyman suggests that "one way to prevent response biases from distorting the hit rate is to use a randomizing procedure that makes sure that each item within a target pool occurs equally often." Coming from a critic as sophisticated as Hyman, this is a very puzzling suggestion, because he appears to be suggesting some variant of sampling without replacement, a procedure that would virtually guarantee response-bias artifacts. For example, if receivers tend to avoid selecting targets that appeared in previous sessions, this response bias would coincide with the actual diminishing probabilities that a previously seen target would reappear. The experimenters—who participate in many sessions and discuss them with one another—are in an even better position to detect and possibly to exploit the diminishing probabilities of target repetition. Sampling without replacement is precisely what enables card counters to improve their odds at blackjack.

Alternatively, perhaps Hyman is advocating a procedure in which the experiment continues until each clip within a judging set appears as a target a predesignated minimum number of times. For purposes of analysis, the investigator then randomly discards excess sessions until the target frequencies are equalized at that minimum number. This would solve the response-bias problem but would be enormously wasteful. Suppose, for example, that only 4 sessions from each judging set would have to be discarded, on average, to equalize the target frequencies. With 40 judging sets, the investigator would end up discarding 160 sessions, equal to nearly half of the sessions that took Honorton and his colleagues 6½ years to collect! Only a study with many fewer judging sets could reasonably implement this strategy.

Hit Rates as a Function of Target Repetition

In his post hoc excursion through the autoganzfeld data, Hyman uncovered an unexpected positive relationship between hit rates and the number of times targets had been targets in previous sessions. (Ironically, Hyman has been one of the most outspoken critics of parapsychologists who search through their data without specific hypotheses and then emerge with unexpected "findings.")

If this finding is reliable and not just a fluke of post hoc exploration, then it is difficult to interpret because target repetition is confounded with the chronological sequence of sessions: Higher repetitions of a target necessarily occur later in the sequence than lower repetitions. In turn, the chronological sequence of sessions is confounded with several other variables, including more experienced experimenters, more "talented" receivers (e.g., Juilliard students and receivers being retested because of earlier successes), and methodological refinements introduced in the course of the program in an effort to enhance psi performance (e.g., experimenter "prompting").

Again, however, Hyman's major concern is that this pattern might reflect an interaction between inadequate target randomization and possible response biases on the part of those receivers or experimenters who encounter the same judging set more than once. This seems highly unlikely. In the entire database, only 8 subjects saw the same judging set twice, and none of them performed better on the repetition than on the initial session. Similar arithmetic applies to experimenters: On average, each of the eight experimenters encountered a given judging set only 1.03 times. The worst case is an experimenter who encountered the same judging set 6 times over the 6½ years of the program. These six sessions yielded three hits, two of them in the first two sessions.

At the end of his discussion, Hyman wonders whether this relationship between target repetition and hit rates is "due to an artifact or [does it] point to some new, hitherto unrecognized property of psi?" If it should turn out to be the latter, then I believe it only appropriate that parapsychologists reward his serendipity by calling it the Hyman Effect.

References

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

Honorton, C., Berger, R. E., Varvoglis, M. P., Quant, M., Derr, P., Schechter, E. I., & Ferrari, D. C. (1990). Psi communication in the ganzfeld: Experiments with an automated testing system and a comparison with a meta-analysis of earlier studies. *Journal of Parapsychology, 54,* 99–139.

Hyman, R. (1994). Anomaly or artifact? Comments on Bem and Honorton. *Psychological Bulletin, 115,* 19–24.

Received August 9, 1993
Accepted August 9, 1993 ■