Discussion paper: Should ganzfeld research continue to be crucial in the search for a replicable psi effect?

By Julie Milton

Abstract: A group of recent, well-controlled ganzfeld studies failed to replicate the positive findings of earlier work (Milton & Wiseman, in press, a). This presents a challenge to claims that a ganzfeld psi effect can be replicated across experimenters under methodologically stringent conditions. Because of the ganzfeld's history as a focus for proof-oriented questions, this situation has implications for parapsychology as a whole. In this paper it is shown that replication of effect size in the recent ganzfeld studies is not demonstated across experimenters regardless of whether the database is updated to include recent studies or whether different outcome and cumulation statistics are applied instead of those preplanned. Problems with interpreting other parapsychological meta-analyses of less clearly well-conducted studies and apparently consistent process-oriented findings as strong evidence for psi are discussed. The case is made for continuing with ganzfeld research as an important focus of parapsychology's claims for replicability. It is argued that if there is a replicable ganzfeld psi effect, however, the procedures necessary to produce it have not yet been identified. It is proposed that process-oriented work be directed to the goal of identifying which studies should be able to replicate an above-chance effect and that these studies, identified by their planned procedures before they have been conducted, should provide the basis for future tests of replication.

Introduction

Despite the field's long history, there is still controversy over whether the results of parapsychology experiments offer evidence for a genuine communication anomaly — "psi". For some time, parapsychologists have recognised that the evidence for psi most likely to convince fair-minded but critical scientists would be an experimental procedure that a range of experimenters could carry out that would produce reasonably replicable effects. Unless the experiment's effects could be replicated across experimenters, there would always remain fraud, error or sensory leakage as strong alternative explanations to the psi hypothesis.

For many years, such replicability appeared to be out of reach. However, this perception appeared to change with the arrival of several research programs involving free-response ESP in the 1970s. Ganzfeld ESP studies in particular seemed especially promising. Not only did a range of experimenters appear to obtain outcomes in ganzfeld studies that were above chance, but they did so under conditions that appeared to be well-controlled and without using specially selected participants. In 1981, Ray Hyman, a psychologist skeptical of the existence of psi, wanted to conduct a critical assessment of a research program that represented parapsychology's strongest evidence. The claims being made for ganzfeld research made it an obvious choice for his attention (Hyman, 1985).

Hyman (1985) meta-analysed the 42 studies conducted since publication of the first ganzeld ESP study in 1974, finding an overall statistically significant outcome. However, he concluded that the methodological problems that he identified in the studies could account for the positive results. In response, Charles Honorton, a proponent of ganzfeld research, conducted his own meta-analysis of the database, restricting his attention to the 28 studies reporting direct hits as an outcome measure (Honorton, 1985). He also obtained a statistically significant overall outcome (see Table 1) but although he conceded that the studies contained potential methodological problems, did not agree that the problems were sufficient to account for the overall outcome.

Rather than continue to dispute the matter, Hyman and Honorton (1986) instead jointly drew up a set of methodological guidelines for the stringent conduct of future ganzfeld studies, agreeing that the case for psi in the ganzfeld would rely on a broad range of experimenters obtaining positive results under such conditions. Meanwhile, Honorton and his research team at Princeton Research Laboratories (PRL) had begun in 1982 a series of partially automated ganzfeld studies — "autoganzfeld studies" — designed to meet Hyman's methodological concerns (Bem & Honorton, 1994; Honorton et al., 1990). Eleven series were completed before PRL closed in 1989, obtaining a statistically significant overall outcome and a mean effect size nearly identical to that obtained in Honorton's (1985) meta-analysis of the earlier ganzfeld database (see Table 1). Replication of the early ganzfeld results under stringent conditions appeared to suggest that methodological problems were unlikely to have entirely accounted for the effects obtained in the earlier studies. However, Bem and Honorton pointed out that it still remained for their results to be replicated by other experimenters under similarly stringent conditions.

In early 1997, Richard Wiseman and I therefore attempted to determine whether other experimenters had indeed succeeded in replicating these results under well-controlled conditions, by meta-analysing the 30 published ganzfeld studies conducted since the publication of Hyman and Honorton's methodological guidelines (Milton & Wiseman, in press, a). The studies' combined outcome was not statistically significant and the mean effect size was near zero (see Table 1). The mean effect size in the recent studies is less than a seventeenth of that found in the PRL work and a post-hoc comparison shows that it is statistically significantly lower than the mean effect sizes of the PRL and earlier ganzfeld databases (see Table A2).

Table 1
Outcomes of meta-analyses of ESP ganzfeld studies.

Meta-analysis	N studies	N trials	Stouff. z	p (1-t)	Effect size[a] Mean	s.d.	95% confidence interval
Honorton (1985)[b]	28	835	6.60	2.2×10^{-11}	0.26	0.38	0.12 to 0.40
Bem & Honorton (1994)	11	329	3.41	.00033	0.23	0.24	0.09 to 0.37
Milton & Wiseman (in press, a)	30	1198	0.70	.24	0.013	0.23	-0.07 to 0.10
All studies 1987 to present[c]	39	1588	2.28	.011	0.038	0.26	-0.04 to 0.12
All studies 1987 to present excl. Dalton (1997a)[c]	38	1460	1.45	.074	0.027	0.25	-0.05 to 0.11

[a]Effect size is $z/N^{1/2}$.
[b]Honorton's meta-analysis solely represents the early ganzfeld database here because Hyman's (1985) report does not provide the number of trials in each study needed for the calculation of $z/N^{1/2}$, the effect size used in this table.
[c]Individual study outcomes were calculated following the same procedures as in Milton and Wiseman (in press, a).

Updating our meta-analysis to include the studies (see Table A1) published to date (March 1999) since our meta-analysis was completed in February 1997 renders the overall cumulation statistically significant[1], but fails to raise the mean effect size to even a sixth of that obtained in the PRL or earlier ganzfeld studies meta-analysed by Hyman (1985) and Honorton (1985) (see Table 1). Moreover, the statistical significance of the updated cumulation is due solely to the inclusion of an extremely successful study by Dalton (1997a) (see Table 1) and not to renewed success by a range of investigators. Whether Dalton's study is included or not, it is clear that the effect size obtained in

Honorton's autoganzfeld studies and in the earlier ganzfeld database has not replicated. Post-hoc comparisons show that the updated database of recent studies, with or without the Dalton study, has a mean effect size statistically signficantly lower than those of the earlier meta-analyses (see Table A2).

The same is true if a variety of alternative outcome calculation and cumulation methods are used to analyse the recent studies rather than the ones that we preplanned and applied (Milton & Wiseman, 1997a). Since the presentation of our meta-analysis at the 1997 Parapsychological Association Annual Convention, a number of colleagues have informally pointed out that a number of different ways of calculating or cumulating individual study outcomes, or the introduction of various criteria for excluding outliers, result in overall statistical significance of varying degrees for the database. Regardless of arguments over the post-hoc and possibly selective nature of these analyses, none of them have the effect of raising the mean effect size in the new database by any meaningful amount, because of the relative insensitivity of means compared to the statistical significance of cumulations when slight changes are made in the treatment of a database. For example, summing the number of direct hits obtained across studies using Bem and Honorton's (1994) method (approximating the number of direct hits from the standard normal deviate of the study's reported outcome measure if direct hits were not reported) results in a total of 331 hits in the 1198 trials in the database. This is a statistically significant outcome ($p = .019$, one-tailed) but the effect size measured in this way is only 0.060.

Implications of the current situation

The current situation, then, is that the studies that appear to form the group proposed by Bem and Honorton (1994) to form a crucial test of the evidence for psi in the ganzfeld have clearly failed to show replication of an above-chance effect across experimenters and only show overall statistical significance to date if one extremely successful study is included. On the face of it, this appears to be an important replication failure because the unique history of ganzfeld research — strong claim, critical assessment, methodological guidelines, methodological refinement, initial replication —have led to it being presented to mainstream science as a critical test of the evidence for psi.

However, it is almost 20 years since Hyman's (1985) meta-analysis placed the focus for assessing the evidence for psi on ganzfeld research. Since that time, meta-analyses have been conducted of other parapsychological databases, including some whose main purpose has been to examine process-oriented hypotheses. The studies within them are not as well-controlled as the recent ganzfeld studies appear to be, but their highly statistically significant cumulated outcomes, their apparent resistance to explanations in terms of selective reporting, their general lack of statistically significant correlations between individual studies' quality and effect size in these databases and the apparent replicability of successful studies within them across experimenters has led to their being presented both within and outside parapsychology as providing strong evidence that psi is a genuine communication anomaly that replicates across experimenters (e.g. Honorton & Ferrari, 1989; Radin, 1997; Radin & Nelson, 1989; Radin & Ferrari, 1991; Utts, 1991).

If they do indeed constitute strong evidence, then the replication failure of the recent ganzfeld studies requires no negative reassessment of the claims for psi nor any action to continue to seek evidence for across-experimenter replication of a psi effect under stringent conditions in the ganzfeld.

Problems in interpreting meta-analyses of studies of uncertain quality

However, even if internal analyses reveal no obvious problems, there are difficulties in interpreting meta-analyses as strong evidence for a phenomenon if the studies they contain are of uncertain or low methodological quality. As can be seen in Table 2, the parapsychological databases examined so far consist of exactly such studies. The table summarises the methodological quality observed in the major parapsychology databases meta-analysed so far that have included individual study quality assessments. Setting aside Honorton's (1985) and Hyman's (1985) quality assessments of the early ganzfeld work, which present some problems of interpretation (see footnotes to the table), it can be seen that in fully half of the databases that reported mean study quality, studies scored on average fewer than half of the available methodological quality points. Only the 14-study free-response sub-database in Honorton et al.'s (1998) meta-analysis contained studies that scored more than two-thirds of the available quality points and it can be argued that important quality criteria were omitted from that quality assessment, such as the prespecification of sample size, the use of blind mentation transcription, the prevention of cues to judges from judging trials out of order and so on (see Milton, 1997). Two meta-analyses did not report mean study quality at all.

The lack of evidence that these databases in general consist of high quality studies introduces the possibility that their outcomes may have been inflated and at worst, entirely caused by methodological flaws. In order to be a matter for concern in parapsychology databases, the effect sizes due to methodological flaws would have to be at least as large as the observed effect sizes and the flaws would have to be present in sufficient quantities (singly or in combination) to be relevant. However, there has been very little empirical research to determine the effect sizes associated with the absence of the various methodological safeguards used in parapsychology (Milton & Wiseman, 1997b) and many meta-analyses do not report the frequency with which individual safeguards are not reported. It is therefore difficult to rule out methodological problems as an explanation for the observed results. There are, in fact, meta-analyses in which flaws likely to be associated with effect sizes not much, if any smaller than those observed appear to be potentially prevalent. For example, not prespecifying which of several possible measures (such as direct hits, ranks, etc. in free-response ESP studies) is to be used to test the null hypothesis clearly has the potential to inflate study outcomes considerably due to post-hoc data selection. The effect size associated with such selection has not been calculated. However, a computer simulation by Hyman (1985) of the effects of being free to choose any of the four main outcome measures available when target ratings are used suggests that the probability of any one of them being statistically significant with an alpha of .05 is approximately .15. In a database of 78 free-response studies (Milton, 1997), the observed probability of a study being statistically significantly above chance was .22 and 96% of studies did not report whether

Table 2

Mean methodological quality of studies in parapsychology meta-analyses expressed as a percentage of the maximum number of quality points available.

Meta-analysis	Effect examined	Mean quality (%)
Honorton (1985)	Ganzfeld ESP	70[a]
Hyman (1985)	Ganzfeld ESP	44[b]
Honorton & Ferrari (1989)	Forced-choice precognition	41
Honorton et al. (1998)	ESP-extraversion relationship:	
	forced-choice studies:	45
	free-response studies:	86
Lawrence (1993)	ESP-belief in psi relationship	46
Milton (1997)	Non-ASC free-response ESP:	
	GESP studies:[c]	61
	clairvoyance studies:[c]	58
	precognition studies:[c]	47
Radin & Ferrari (1991)	Dice PK	not reported
Radin & Nelson (1989)	Micro-PK	not reported
Stanford & Stein (1994)	ESP-Hypnosis relationship	49
Steinkamp et al. (1998)	Precognition vs clairvoyance:	
	clairvoyance studies:	66
	precognition studies:	63

Note: The meta-analyses used different quality criteria, ranging from 2 to 18 safeguards being examined in each meta-analysis. The mean quality of each meta-analysis is therefore not directly comparable with another.

[a]In this meta-analysis, Honorton assessed study quality on just two features — the availability of sensory cues from target handling and the adequacy of the target randomisation method. He assigned partial credit to studies containing methodological features (the use of single rather than duplicate target sets and randomisation using hand shuffling, coin-flipping or die-throwing) that have received no credit in other parapsychological meta-analyses (Honorton & Ferrari, 1989; Lawrence, 1993; Milton, 1997; etc.). This method allowed him to make a distinction between these studies and studies using less stringent or unknown methods but for the purposes of this table arguably inflates apparent study quality by a considerable amount. For example, all but one study received at least one quality point for preventing sensory cueing regardless of whether a duplicate target set was used. If quality points are assigned in a manner more consistent with the other meta-analyses, with one point for the use of duplicate judging sets and no points for manual methods of randomisation, the studies obtained 46% of the maximum available quality points.

[b]Based on only 4 of Hyman's 12 flaw categories. One of the excluded categories involved assigning a flaw to studies in which it was not clear that receivers' friends were used as senders. This does not seem appropriate because it is absence of appropriate security rather than the relationship between participants that would constitute an inadequate precaution against collusion. The remaining 7 flaws concerned statistical errors and the use of multiple outcome measures without adjustment for multiple analysis. They could not have affected study outcomes in the meta-analysis because Hyman calculated outcomes using appropriate statistics and single measures and are not therefore included here.

[c]The original paper reports these percentages in terms of publication type rather than study type.

the choice of outcome measure was preplanned. Hyman's study is likely to provide an extreme upper limit for the action of this particular flaw since it is not probable that post-hoc selection of statistically significant outcome measures happens in every study, as it did in his simulation. Nevertheless, the potential effects of not prespecifying outcome measures is clearly not trivial in comparison with the outcomes of ESP studies. Similarly, recording errors have been estimated empirically to occur on approximately 1% of trials and to be biased in favour of the observer's hypothesis on two-thirds of the trials (Rosenthal, 1978). The mean effect size in Honorton and Ferrari's (1989) database of forced-choice precognition studies is equivalent to raising a study's outcome 1% above a mean chance expectation of 50% but the frequency with which studies reported double-blind, double-checked or automated data recording is not reported.

In most parapsychological meta-analyses, estimates of overall study quality do not correlate statistically significantly with effect size. A number of the researchers who obtained such null correlations have concluded that methodological problems therfore had no meaningful influence on their databases (e.g., Honorton & Ferrari, 1989; Lawrence, 1993; Radin & Ferrari, 1991; Radin & Nelson, 1989). However, in databases that do not consist entirely or mostly of clearly well-controlled studies such as the parapsychology databases, there are many ways in which a relationship between methodological flaws and effect size could be obscured. This is a general problem in meta-analysis and not one restricted to parapsychology. However, these problems have received little attention in parapsychology (although see Hyman, 1985; Milton, 1997; Stanford & Stein, 1994) and so it is worth listing some of them. A selection, by no means exhaustive, is as follows:

(1) The absence of safeguards for certain procedures (such as randomisation or sensory shielding procedures) might inflate effect size than others (such as lack of double-blind checking of data records). In an unweighted correlation of study quality and effect size, the effect of the absence of these more important safeguards might be drowned out by the other data (Stanford & Stein, 1994). In some cases, experts have been called upon to rate flaws in terms of their likely impact so that a weighted correlation can be performed between the absence of safeguards and effect size (e.g., Milton, 1997; Radin & Ferrari, 1991). Thus far, these weightings have not indicated any such relationships, but it could be argued that, given the general lack of direct empirical evidence concerning effect sizes that result from the absence of safeguards, the experts' judgements may be wrong, regardless of how well they agree with each other.

(2) It is unlikely that individual studies' methodological quality is accurately reflected by their quality coding. Most parapsychological studies, especially those conducted before the 1980s, have not been written with a future meta-analyst's quality checklist in mind and it is often unclear from reports whether particular safeguards against sensory leakage, error, post-hoc data selection and so on have been carried out. Presented with unclear or circumstantial evidence concerning the presence of a safeguard, coders will have to make a subjective judgement according to this partial, ambiguous information, influenced by their individual expectations and assumptions about what experimenters are likely to do

as a matter of standard laboratory procedure. Under these circumstances, errors in coding are very likely to arise.

(3) The binary coding of methodological safeguards as either present of absent in almost all parapsychology meta-analyses to date means that studies whose use of the safeguard is unknown must be included in the "safeguard present" or "safeguard absent" group. For example, it may be assumed that studies whose reports do not address the issue of study size at all belong with studies that clearly did not prespecify the number of trials to be conducted as a safeguard against optional stopping (Milton & Wiseman, 1997b). However, given that at least some, but by no means all experimenters are likely to have used the safeguard without reporting it, this will result in a group of studies that all used the safeguard being compared with a group of studies in which some used the safeguard and some did not, in an unknown proportion. If the studies that did not use the safeguard had higher effect sizes as a result, then including the studies that used but did not report the safeguard in the same category will reduce the average effect size in that group, bringing it closer to that of the group that clearly used the safeguard. This would clearly reduce the sensitivity of a test comparing the mean effect sizes in the two groups and could obscure a genuine relationship between effect size and methodological quality. The only parapsychological meta-analysis published so far that allowed assessors to code the presence of a safeguard in a study as unknown rather than merely present or absent found that up to 59% of studies fell into this category on certain safeguards (Steinkamp et al., 1998), suggesting that the problem may be by no means trivial in other parapsychology databases.

(4) The binary quality ratings used in parapsychological meta-analyses may also lead to insensitive quality analyses because they are crude measures of quality, whereas the seriousness of a flaw may often vary more smoothly in magnitude than this. For example, the use of card-shuffling to randomise the target sequence in an ESP study would count as a flaw in most parapsychological meta-analyses, but, because randomness improves as the number of shuffles increases, a study in which the deck was shuffled a lot would be less prone to error than a study in which the deck was only shuffled a few times. Analyses based on the usual binary flaw ratings may be too insensitive to pick up a relationship between flaws and effect size (Stanford & Stein, 1994).

(5) Experimenters who obtain null results in their studies may give shorter accounts of them that leave out details of the safeguards that they included (as Pratt, 1966, states that he did, for example). In a meta-analysis, such studies as a group would show a spurious association between low effect size and low quality that might act to hide a real association between low effect size and high quality in the other studies in the database (Milton, 1997).

(6) Quality coding has almost always been conducted non-blind in parapsychology meta-analyses so that it is difficult to rule out the possibility of coders being influenced in their coding by the studies' outcome. Coders who favour the psi hypothesis might be reluctant to ascribe flaws to successful studies or, conversely, might overcompensate for

their bias by being more ready to penalise unsuccessful studies. Either strategy would introduce error variance.

(7) Flaws might not behave additively, but might instead interact with each other, reducing the sensitivity of simple contrast or correlation analyses that examine the relationship between total flaws and effect size (Stanford & Stein, 1994). Similarly, the relationship between the lack of any given safeguard and effect size might not be linear; a flaw may only become 'active' above a certain threshold, for example, and again, a simple correlative approach would be insensitive to this (Stanford & Stein, 1994).

(8) If the presence of some flaws is negatively correlated, they might raise effect sizes in the database but their effects would be difficult to detect. A database in which either safeguard A or safeguard B is present in each study, but never both together, serves as an extreme example to illustrate the point. If the absence of each safeguard increases effect size to roughly the same degree, then a comparison of effect sizes of studies in which safeguard A is present with studies in which it is absent will show no difference; nor will such a comparison show any difference when applied to safeguard B (Hyman, 1985).

There are plenty of reasons, then, for being cautious about concluding that methodological flaws do not increase study outcomes because estimates of studies' overall methodological quality are not statistically significantly correlated with effect size. Moreover, if the effects of flaws cannot be ruled out then the other aspects of the meta-analyses' results that appear to support the psi hypothesis — that is, implausibly large "filedrawers" of unpublished, null studies required to render the overall cumulation non-signficant, and replicability across experimenters — also are undermined. If study outcomes in a meta-analysis have been inflated by flaws then so has the size of the "filedrawer". If it were possible to correct study outcomes for the influence of those flaws, the overall cumulation will fall and the filedrawer may no longer appear unreasonably large. Concerning replicability, all of the parapsychological databases that have examined it have shown stastistically significant heterogeneity of effect size across studies (Honorton & Ferrari, 1989; Honorton, Ferrari & Bem, 1998; Milton, 1997; Radin & Ferrari, 1991; Radin & Nelson, 1989; Stanford & Stein, 1994) with the exception of the PRL ganzfeld database (Honorton et al., 1990). The replicability that many of them claim is replicability of successful rejection of the null hypothesis, using a variety of methods. Honorton and Ferrari (1989), for example, report that 30% of studies and 37% of experimenters obtained statistically significant results, indicating that more sucessful outcomes were obtained than the 5% expected by chance and that success was not restricted to a few experimenters. Clearly, replicability defined in these terms is also vulnerable to explanation in terms of methodological artefacts in databases in which quality is unclear.

There is, however, a second type of evidence for psi that is often mentioned in addition to the results of proof-oriented meta-analyses, and that is that a number of literature reviews and meta-analyses of process-oriented psi research appear to indicate consistent relationships between study outcomes and variables such as belief in psi, extraversion

and so on. It appears to be often assumed that such relationships would not consistent if they were attributable to methodological flaws. For example, it may be assumed that the "sheep-goat" effect in which believers in psi score higher on psi tasks than non-believers cannot be due to sensory cues because these cues would be equally available to both sheep and goats and both groups would be expected to show the same level of performance.

This is not a safe assumption, however. There are many situations in which one might expect the action of flaws to produce consistent differences between groups in line with parapsychologists' hypotheses. For example, in sheep-goat studies that do not have adequate sensory shielding, participants might be expected to be motivated to exploit those sensory cues (consciously or otherwise) to perform in accordance with their beliefs, just as they are hypothesised to do with extrasensory cues — believers to score more hits and goats to score fewer hits. The pattern of the results due to the inadequate sensory shielding would mimic that expected under the usual sheep-goat hypothesis. As Palmer (1978) notes, the results of ESP experiments tend to fall into patterns that make psychological sense, inasmuch as they appear similar to the patterns of results that one might expect if subjects were attempting to respond to very weak sensory information. Many spurious results due to flaws would also be expected to make similar sense, however, especially if they were in fact based on sensory leakage (see also Wiseman & Morris, 1995). Many of the more consistent findings of ESP research, such as higher scoring on confidence calls than on other trials (Carpenter, 1977; Palmer, 1978), higher scoring in studies with trial-by-trial feedback (Honorton & Ferrari, 1989), and so on, make conventional psychological sense if one assumes that they are due to the exploitation of weaknesses in the design by participants. Psychologically meaningful and consistent patterns of results would also be expected if safeguards preventing experimenter bias were lacking, such as predetermination of study sizes, prespecification of statistical tests, data checking and so on. Arguing that process-oriented research has shown a consistent and meaningful pattern of results does not, therefore, allow side-stepping of the question of methodological quality if this argument is to be used in a proof-oriented way. Furthermore, it is difficult to make a strong proof-oriented case on the basis of this process-oriented work because meta-analyses of studies examining relationships between apparent ESP performance and moderator variables indicate similar problems of low or unclear quality in studies as are found in the proof-oriented meta-analyses (Honorton, Ferrari & Bem, 1998; Lawrence, 1993; Stanford & Stein, 1994).

I am not arguing that methodological problems clearly account for the positive results of the parapsychological meta-analyses. The study quality estimates that the meta-analyses report are in most cases minimum estimates of quality because they conservatively do not give the benefit of the doubt to studies that do not report details of safeguards; the actual quality of the studies may have been higher than it appears. The general absence of demonstrable relationships between studies' quality estimates and their effect sizes is encouraging for the psi hypothesis, if not a matter for complacency. Nor is it my intention to discourage the use of meta-analysis as a valuable tool because it cannot

answer all of the questions that we would want to ask about a database. It is clearly a more powerful method for synthesising research findings than traditional literature reviews. However, there appear to be potentially serious problems with drawing strong conclusions from reviews and meta-analyses of studies that are not demonstrably strong in quality and these problems apply as much to process-oriented research as they do to proof-oriented research. If providing strong evidence for psi is still seen as important, then it appears that the only way to do so is by demonstrating a replicable, non-zero effect across a range of experimenters under stringent methodological conditions. So far, this does not appear to have happened.

Implications for future research

Ganzfeld research seems an obvious area in which to continue to look for strong evidence for psi. No other research methodology in parapsychology has received the detailed critical attention that ganzfeld research has received; it is the only area in which a whole database of studies has been examined intensively by a researcher such as Hyman who considers the existence of a genuine anomaly unlikely (Hyman, 1985) and in which researchers with opposing viewpoints have jointly produced methodological guidelines for research to settle the question of the existence of psi (Hyman & Honorton, 1986). In addition, it has arguably come to represent the case for psi in microcosm for mainstream psychology (Bem & Honorton, 1994; Milton & Wiseman, in press, a) and an account of it appears in every major summary of parapsychology's best evidence for psi (e.g., Atkinson et al., 1990; Broughton, 1992, Hayes, 1998; Krippner et al., 1993; Radin, 1997; Utts, 1991). The failure of the recent studies to replicate the success of the earlier work therefore presents a challenge in the same, mainstream scientific forum to parapsychology's claims for a genuine, replicable effect.

If ganzfeld research is to be an important player in the continued search for strong evidence, that search will only be successful if a replicable effect can be demonstrated. At present, however, if there is no change in the way ganzfeld research is carried out and no change in how replicability is examined, there appears to be no obvious reason why the next, inevitable meta-analysis of future ganzfeld studies will not show the same pattern of a null, or near-null cumulation with perhaps a few individual experimenters obtaining effects that others are not replicating. In order to avoid repeating recent history we need to know why the recent meta-analysis (Milton & Wiseman, in press, a) failed to replicate the findings of the PRL studies, which were carried out under similarly stringent conditions.

Unfortunately, the explanation is far from clear. One possible reason could be that the results of earlier ganzfeld studies were due to methodological problems rather than to psi. However, although a number of potential avenues for sensory leakage have been identified in the PRL work (Honorton et al., 1990; Morris et al., 1993; Wisesman et al., 1996), none appear sufficiently strong to account in any obvious way for the success of those studies, which were much more well-controlled than the earlier work (Milton & Wiseman, in press, a).

A second possibility is that the PRL studies used psi-conducive procedures but that the recent studies did not. This is possible but far from certain, for two reasons. First, although Bem and Honorton (1994) identified a number of variables that may be important for replication, the vast majority of recent studies meta-analysed (Milton & Wiseman, in press, a) either did not measure or did not report the average values of these variables in their studies (with the exception of the use of static versus dynamic targets where it is clear that the two databases are closely matched) and so it is not possible to make a strong case that differences in these variables accounted for the lack of replication (Milton & Wiseman, in press, a).

Second, it is possible that any number of additional, unidentified variables might have contributed to the success of the PRL studies and so it is not possible to know whether the recent studies' failure to replicate the PRL work was due to their failure to exploit these variables to the same extent. There were a number of procedures used on all or almost all trials at PRL — the use of a sender, continuous auditory monitoring of the receiver's mentation by the sender, correspondence judging by the receiver rather than by an independent judge, (double-blind) prompting by the experimenter during the judging to correspondences that the receiver overlooked, a 14-minute pre-trial relaxation procedure for both sender and receiver, and so on — whose importance has not been empirically determined. Any one or more of these procedures might be important for replication. However, without any evidence for their effects, it is not clear that the failure of the recent studies to replicate the findings of the PRL studies was due to the use of different procedures. It is not evident, at this point, what a replication of the PRL work in its essentials would have to consist of.

Since the convention presentation of our meta-analysis (Milton & Wiseman, 1997a), a number of colleagues have informally suggested that if we had restricted our database to "standard" ganzfeld studies (that is, studies without unusual features), then across-experimenter replication of the PRL effect size might have been evident. However, there appears to be little agreement among the researchers who have discussed the issue with me on what the features of a standard ganzfeld study would be. Devising a rule to define such a study at this point could easily appear as a post-hoc attempt to explain away a disappointing result, given that the previous ganzfeld meta-analyses included almost all studies and trials no matter how unusual their procedures (Bem & Honorton, 1994; Honorton, 1985; Hyman, 1985) and regardless of whether those procedures would be expected to result in success or failure[2]. Neither Hyman and Honorton (1986) nor Bem and Honorton (1994) specified that studies would have to have certain features to be considered part of the replicability test that they proposed. It does not appear possible to selectively meta-analyse the recent studies and make a strong case that the ganzeld effect is replicable.

However, a selective meta-analysis with exclusion criteria stated in advance of studies being conducted would be a credible demonstration of replicability if it obtained positive results. In practice, it is unlikely that criteria could be set up that would anticipate all of the novel features that experimenters might introduce in their studies that would lead

most researchers to expect them to be unsuccessful. In addition to having to conform to a basic set of criteria, the procedures planned for each study would therefore also have to be examined on a case-by-case basis to determine whether the study ought to be included in the replication test or not. The existence of such a project would not affect the usual conduct of process-oriented research or force experimenters to use certain procedures in their studies. It would simply be the case that studies eligible to be included in the meta-analysis would be included and others would not. Similarly, the project would not affect anyone's usual freedom to conduct a meta-analysis of their own. In particular, there is no reason why anyone should not conduct a process-oriented meta-analysis involving all studies.

Some researchers may believe that it is already possible to identify successful ganzfeld studies based on their procedures alone and that it would be advisable to begin such a meta-analysis now. Others may think this premature. Very few variables have been explored repeatedly or systematically in ganzfeld studies and even fewer have been examined meta-analytically across studies to determine whether there is good statistical evidence that they relate to effect size. Meta-analytic investigation of some of the variables suggested by Bem and Honorton (1994) as having been important in the PRL work indicates that other experimenters have not replicated their effects in the few areas where this has been attempted (Milton & Wiseman, in press, a). In addition, some variables identified by Bem and Honorton as having had statistically significant relationships with effect size in the PRL studies do not in fact appear to have done so (Milton & Wiseman, in press, a), suggesting that our success so far in identifying what variables are important in the ganzfeld might be more limited than has been assumed. It may be that a systematic assessment of process-oriented ganzfeld research is called for before embarking upon a replication test that should exploit its findings (see, e.g., Dalton, 1997b).

Summary and conclusion

The meta-analysis of recent, well-controlled ganzfeld studies (Milton & Wiseman, in press, a) indicates a failure to replicate the results of the earlier work, and the evidence for psi from meta-analyses and process-oriented reviews of parapsychology studies of low or uncertain quality does not appear compelling. If the search for strong evidence for psi is to continue, ganzfeld research appears to be its natural arena. A meta-analysis that excludes studies before they are conducted if they are not expected to replicate a positive effect appears to be the obvious test of future replication. Such a meta-analysis may be premature until more research has been done to identify what factors may be psi-conducive in the ganzfeld but it appear to be an important goal to work towards.

Many researchers may disagree with my assessment of the evidence for psi accumulated so far and with my goal of continuing to seek stronger evidence in general and my proposal for a prospective ganzfeld meta-analysis in particular. Conversely, many may disagree with the use of meta-analyses of studies of uncertain quality being promoted as strong evidence for psi and with ganzfeld research having become a crucial test-case before the factors that affect its replicability have been well-established. Whatever

researchers' views may be, however, the momentum of previous events is carrying the field towards another inclusive meta-analysis of future ganzeld studies that appears likely to show the same failure to replicate as did the last one. A second failure to replicate that occurs despite the warning of a first failure will give the appearance of reasonably strong evidence against claims for psi as a replicable (and therefore, probably genuine) effect.

If this is not a direction that parapsychologists want events to take, then now appears to be the time to say so. Although the choice of whether to carry out a meta-analysis is likely to be an individual one, its results will affect other researchers. The opportunity for the research community, rather than a few, key individuals, to discuss the issues and express their opinions is long overdue and I look forward to hearing the views of my colleagues on the matters that I have discussed in this paper.

References

Atkinson, R.L., Atkinson, R.C., Smith, E.E., & Bem, D.J. (1990). *Introduction to psychology (Tenth Edition).* Orlando, FL: Harcourt Brace Jovanovich.

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

Broughton, R. (1992). *Parapsychology: The controversial science.* London: Rider.

Carpenter, J.C. (1977). Intrasubject and subject-agent effects in ESP experiments. In B.B. Wolman (Ed.), *Handbook of parapsychology*, pp. 202-272. New York, NY: Van Nostrand Reinhold.

Dalton, K. (1997a). Exploring the links: Creativity and psi in the ganzfeld. In *The Parapsychological Association 40th Annual Convention proceedings of presented papers*, pp. 119-134. The Parapsychological Association.

Dalton, K. (1997b). Is there a formula to success in the ganzfeld? Observations on predictors of psi-ganzfeld performance. *European Journal of Parapsychology*, 13, 71-82.

Hayes, N. (1998). *Foundations of psychology: An introductory text (Second Edition).* Walton-on-Thames, England: Nelson.

Honorton, C. (1985). Meta-analysis of psi ganzfeld research: A response to Hyman. *Journal of Parapsychology*, 49, 51-91.

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.

Honorton, C. & Ferrari, D.C. (1989). Meta-analysis of forced-choice precognition experiments. *Journal of Parapsychology*, 53, 281-308.

Honorton, C., Ferrari, D.C., & Bem, D.J. (1998). Extraversion and ESP performance: A meta-analysis and a new confirmation. *Journal of Parapsychology*, 62, 255-276.

Hyman, R. (1985). The ganzfeld psi experiment: A critical appraisal. *Journal of Parapsychology*, 49, 3-49.

Hyman, R., & Honorton, C. (1986). A joint communiqué: The psi ganzfeld controversy. *Journal of Parapsychology*, 50, 350-364.

Krippner, S., Braud, W., Child, I.L., Palmer, J., Rao, K.R., Schlitz, M., White, R.A., & Utts, J. (1993). Demonstration research and meta-analysis in parapsychology. *Journal of Parapsychology*, 57, 275-286.

Lawrence, T. (1993). Gathering in the sheep and goats... A meta-analysis of forced-choice sheep-goat ESP studies, 1947-1993. In *The Parapsychological Association 36th Annual Convention: Proceedings of presented papers,* pp. 75-86. The Parapsychological Association.

Milton, J. (1997) Meta-analysis of free-response studies without altered states of consciousness. *Journal of Parapsychology*, 61, 279-319.

Milton, J., & Wiseman, R. (1997a) Ganzfeld at the crossroads: A meta-analysis of the new generation of studies. Paper presented at the Parapsychological Association 40th Annual Convention, Brighton, England.

Milton, J., & Wiseman, R. (1997b). *Guidelines for extrasensory perception research.* Hatfield, England: University of Hertfordshire Press.

Milton, J., & Wiseman, R. (in press, a) Does psi exist? Lack of replication of an anomalous process of information transfer. *Psychological Bulletin.*

Milton, J., & Wiseman, R. (in press, b) A meta-analysis of mass-media tests of extrasensory perception. *British Journal of Psychology.*

Morris, R.L., Cunningham, S., McAlpine, S., & Taylor, R. (1993). Toward replication and extension of autoganzfeld results. The *Parapsychological Association 36th Annual Convention: Proceedings of presented papers*, pp. 177-191. The Parapsychological Association.

Palmer, J. (1978). Extrasensory perception: Research findings. In S. Krippner (Ed.), *Advances in parapsychological research 2: Extrasensory perception,* pp. 59-243. New York: Plenum Press.

Parker, A., & Westerlund, J. (1998). Current research in giving the ganzfeld an old and a new twist. In *The Parapsychological Association 41st Annual Convention proceedings of presented papers*, pp. 135-142. The Parapsychological Association.

Pratt, J.G. (1966). New ESP tests with Mrs. Gloria Stewart. *Journal of the American Society for Psychical Research*, 60, 321-339.

Raburn, L. (1975). *Expectation and transmission factors in psychic functioning.* Unpublished honors thesis, Tulane University, New Orleans, LA.

Radin, D.I. (1997). *The conscious universe: The scientific truth of psychic phenomena.* New York, NY: HarperCollins.

Radin, D.I., & Ferrari, D.C. (1991). Effects of consciousness on the fall of dice: A meta-analysis. *Journal of Scientific Exploration*, 5, 61-83.

Radin, D.I., & Nelson, R.D. (1989). Evidence for consciousness-related anomalies in random physical systems. *Foundations of Physics*, 19, 1499-1514.

Rogo, D.S., Smith, M., & Terry, J. (1976). The use of short-duration ganzfeld stimulation to facilitate psi-mediated imagery. European Journal of Parapsychology, 1, 72-22.

Rosenthal, R. (1978). How often are our numbers wrong? *American Psychologist*, 33, 1005-1008.

Stanford, R.G., & Stein, A.G. (1994). A meta-analysis of ESP studies contrasting hypnosis and a comparison condition. *Journal of Parapsychology*, 58, 235-269.

Steinkamp, F., Milton, J., & Morris, R.L. (1998) A meta-analysis of forced-choice experiments comparing clairvoyance and precognition. *Journal of Parapsychology*, 62, 193-218.

Symmons, C., & Morris, R.L. (1997). Drumming at seven Hz and automated ganzfeld performance. In *The Parapsychological Association 40th Annual Convention proceedings of presented papers*, pp. 441-453. The Parapsychological Association.

Utts, J. (1991). Replication and meta-analysis in parapsychology. *Statistical Science*, 6, 363-403.

Wezelman, R., & Bierman, D.J. (1997). Process oriented ganzfeld research in Amsterdam Series IV B (1995): emotionality of target material, Series V (1996) and Series VI (1997): judging procedure and altered states of consciousness. *In The*

Parapsychological Association 40th Annual Convention proceedings of presented papers, pp. 441-453. The Parapsychological Association.

Wezelman, R., Gerding, J.L.F., & Verhoeven, I. (1997). Eigensender ganzfeld psi: An experiment in practical philosophy. *European Journal of Parapsychology*, 13, 28-39.

Wiseman, R., & Morris, R.L. (1995). Guidelines for testing psychic claimants. Amherst, NY: Prometheus.

Wiseman, R., Smith, .D., & Kornbrot, D. (1996). Exploring possible sender-to-experimenter acoustic leakage in the PRL autoganzfeld experiments. *Journal of Parapsychology*, 60, 97-128.

Department of Psychology
University of Edinburgh
7 George Square
Edinburgh EH8 9JZ
Scotland, UK
101335.2551@Compuserve.com

Appendix

Table A1

Ganzfeld studies published to date (March 1999) since completion of Milton & Wiseman (in press, a) meta-analysis (February 1997).

Study (N = 12)	N trials	z	$z/N^{1/2}$
Dalton (1997a)	128	5.26	0.46
Parker & Westerlund (1998) Study IV	30	2.40	0.44
Parker & Westerlund (1998) Study V	30	1.25	0.23
Parker & Westerlund (1998) Serial Ganzfeld	30	-[a]	-[a]
Symmons & Morris (1997) Pilot Study	12	-[b,c]	-[b,c]
Symmons & Morris (1997) Main Study	51	2.98[b]	0.42[b]
Wezelman & Bierman (1997) Amsterdam Series IV B	32	-1.48	-0.26
Wezelman & Bierman (1997) Amsterdam Series V	40	-0.91	-0.14
Wezelman & Bierman (1997) Amsterdam Series VI	40	-0.15	-0.02
Wezelman & Bierman (1997) Amsterdam Series VI Exploratory Meditation Trials	7	-1.11	-0.42
Wezelman & Bierman (1997) Amsterdam Series VI Exploratory Psilocybine Trials	12	-[d]	-[d]
Wezelman et al. (1997)	32	2.15	0.38

[a]In this study, the receiver's task was to place the four targets in the judging set in the order in which they had been presented during the ganzfeld session. The authors present the results as a frequency table of the number of correct placements within each trial. By inspection the outcome is slightly below chance. However, the authors do not present or refer to any specific inferential statistical analysis and because it is not clear what analysis was intended, no post-hoc analysis has been imposed here.

[b]In both studies by Symmons and Morris, tapes of drumming at different frequencies were used instead of white noise and so it is questionable whether they can be considered as using a ganzfeld environment. The studies are included here to make clear the effects on the database of including or excluding them.

[c]No outcome was reported for the pilot trials.

[d]Two receivers guessed at the same target on 6 trials, obtaining 7 hits in the resulting 12 trials. However, data are not presented that would allow for correction of the non-independence of their calls (the "stacking effect": see Milton & Wiseman, in press, b) and so no outcome is presented here.

Table A2
Post-hoc comparisons between mean effect sizes in meta-analyses of recent and earlier ganzfeld studies.

Databases compared	***t***	**d.f.**	***p*(one-tailed)**
Honorton (1985) vs.:			
Milton & Wiseman (in press, a)	3.06	56	.0017
All studies 1987 to present	2.90	65	.0026
All studies 1987 to present excl. Dalton (1997a)	3.04	64	.0017
Bem & Honorton (1994) vs:			
Milton & Wiseman (in press, a)	2.64	39	.0059
All studies 1987 to present	2.22	48	.016
All studies 1987 to present excl. Dalton (1997a)	2.38	47	.011

Footnotes

[1]It was not possible to calculate outcomes for three of the studies (see footnotes to Table A1) but given that one of these studies (Parker & Westerlund, 1998, Serial Ganzfeld) is clearly slightly below chance and the remaining two studies are very small with only 12 trials each, it is unlikely that their results would increase the cumulated outcome of the database by a meaningful amount.

[2]The previous ganzfeld meta-analyses did not report explicit exclusion rules but the implicit rules appear to have been to include every ganzfeld study (for Hyman's meta-analysis) or every single trial (for the PRL meta-analysis) in which a ganzfeld environment (even a modified one) was used to conduct an ESP test, with one disputed exception. For the first meta-analysis of ganzfeld studies, Honorton provided Hyman with "a copy of every ganzfeld study known to him" (Hyman, 1985, p. 4), all of which Hyman included in his meta-analysis. The studies were procedurally very varied, with some having features that laboratory lore might predict would not be psi-conducive, such as very short mentation periods (e.g. Rogo et al., 1976). However, Honorton did exclude two conditions in a study by Raburn (1975) in which participants were not aware that they were taking part in an ESP test, on the grounds that these trials were too atypical of other ganzfeld research. Hyman (1985) objected to their exclusion because other studies contained unique features and yet were included in the database. Bem and Honorton's (1994) subsequent meta-analysis of the PRL work included every single trial done using the autoganzfeld. The PRL studies were also procedurally varied and the meta-analysis included trials that, again, might arguably not be expected to be successful, such as demonstration trials carried out in the presence of a TV crew and trials from Series 302 in which Target 79 was included in the target set on each trial despite its never having been previously correctly identified when serving as the target.

Acknowledgements

The writing of this paper was generously supported by the Fundacao Bial and the Society for Psychical Research. I am grateful to Bob Morris for helpful comments on an earlier draft.