A SURVEY OF FREE-RESPONSE JUDGING PRACTICES

Julie Milton
Psychology Department
University of Edinburgh
7 George Square
Edinburgh EH8 9JZ
Scotland, U.K.

ABSTRACT

An idealised model of the free-response judging process is developed, and its elements discussed in terms of judging practices in those free-response studies published in full between 1964 and 1985. A wide variety of occasionally conflicting judging practices was found, along with valuable indications for further research in this important area.

ACKNOWLEDGEMENTS: My thanks are due to Nancy Zingrone and Deborah Weiner for allowing me to use a draft version of their free-response bibliography.

While free-response methodology has been popular in ESP studies over recent years, very little research has been directed to the important question of how best to judge the correspondence between free-response material and the target. However, many experimenters have commented on judging issues, or have reported relevant analyses or data which, when brought together, may suggest strengths and weaknesses in our judging practices, and promising directions for future research.

With these aims in mind, I have examined various aspects of procedure which might influence the success of judging, using as a database eighty-five free-response studies in which statistical assessment of the results was attempted and which were published in full between 1964 and 1987 inclusive, in the Journal of Parapsychology, Journal of the American Society for Psychical Research, Journal of the Society for Psychical Research, International Journal of Parapsychology, and European Journal of Parapsychology. Space constraints prevent me from presenting a summary table of these studies and their full references, but these can be obtained from me on request. All of the papers in these journals (whether experimental or not), and those appearing in Research in Parapsychology during the same period, were searched for commentary relevant to free-response judging, as well as other sources where appropriate. The survey is in two sections. In the first section, a model of an ideal judging process is presented, and its elements discussed in terms of their importance in current judging practices. The second section addresses the issues of whether percipients or independent judges are best suited to perform the complex judging task, and what qualities a judge should have. Finally, the findings of the review are discussed with their implications for further research.

THE ELEMENTS OF JUDGING

The underlying structure of the judging process

In a free-response ESP experiment, the percipient's task is to observe and report his or her thoughts, imagery, feelings and mental or physical experiences, which might relate to a randomly selected target. In free-response studies, the targets used are generally fairly complex (they may be people, or geographical locations, objects, and so on). The targets may have elements (such as colour, the presence or absence of people) which differ in their salience for the percipient, and in their frequency of occurrence. In addition, targets may be regarded as possessing various broad categories of content (such as semantic content, or emotional content), each of which broad categories may differ in their salience. The salience of both individual elements and categories of content may differ from one percipient to another, depending on individual differences.

Just as free-response targets are complex and varied, so too are the mentations reported by percipients. Mentations may be in the form of imagery in any sense modality, or merely abstract concepts; the may be vivid, bizarre, fleeting, spontaneous, or have other distinguishing characteristics. Content of various kinds may be present in them, with varying chance frequencies of occurrence. Mentation items may relate in a variety of ways to the target material, such as semantically or by association, and to a greater or lesser degree. The type of correspondence may vary from percipient to percipient, or from mentation to mentation, or both. Certain types of mentation, and certain kinds of target-mentation correspondence may be more likely to carry psi information than others.

The function of a free-response judge (in process-oriented research at

least) must be [illegible] the ESP [illegible] trial, the probability that psi was responsible for any resemblance between the target and the mentation (or inversely, the strength of the ESP component on a given trial). In the complex situation described above, one way of looking at the task of an ideal judge is that he or she should:

(i) Assign some numerical value in proportion to the degree of correspondence between a single mentation item and the target (and, in some types of judging, to the controls);

(ii) increase this value (given a perfect match) in accordance with the rarity of occurrence of the mentation item's content in the mentation of all percipients in similar experimental conditions (or in the mentation of that particular percipient on other trials, if such data is available);

(iii) increase this value (given a perfect match) in accordance with the rarity of occurrence of the mentation item's content in the entire experimental target pool;

(iv) increase this value in accordance with the likelihood that the mentation item, by virtue of its characteristics, is psi-related (e.g., whether it was bizarre, vivid, spontaneous, or whatever characteristics, if any, are shown by research to mediate ESP)

(v) increase this value in accordance with the salience which the content of the mentation has for the percipient (e.g. if research shows that the presence of people in a target is highly salient to a percipient, then a mentation item bearing on the presence or absence of people would be weighted relatively heavily);

(vi) increase this value in accordance with the likelihood that the type of correspondence (semantic, emotional, etc.) between mentation item and target carries psi-related information, if such differences in likelihood are indicated by research.

Having thus arrived at a weighted measure of the correspondence between each mentation item in a trial and the target (and controls if appropriate), the measures may be summed across the trial or otherwise combined to yield the ESP score for that trial. Although this procedure resembles an atomistic judging procedure most closely in its structure, it can also be thought of as an implicit or idealised basis for holistic or coded judging procedures. In holistic judging, it is possible to think of the overall rating assigned to items in the judging set as a sum of individual mentation ratings weighted as appropriate. In coded judging, the decision of whether a given content category was present or absent could be regarded as being made according to the sum of weighted ratings of relevant mentation items. Further weightings could then be assigned to each decision according to the known salience of the content category and the rarity of that value of the code in the target pool.

<u>The importance of elements of judging in the literature</u>

Each of the six elements of judging in various forms has received occasional attention either implicitly or overtly in experimental and theoretical papers, although very little direct or systematic research has been done on this topic. Most opinion about how best to judge free-response material seems to be based on anecdotal observations. While such observations may be unreliable, they may also contain useful information about aspects of judging which should be investigated empirically. This being so, each of the six elements of judging is discussed in turn below in the context of commentary and experimental results in the literature surveyed.

(i) Assignment of a numerical value to correspondence

Ideally, the value assigned to the correspondence between a mentation item and a target should reflect the correspondence in some objective (and hence reliable) way. 16 studies reported in 10 papers in the database surveyed used atomistic judging, but in no case was interjudge reliability calculated for the allocation of such ratings. In eight of the studies, each point on the rating scale was labelled for the use of the judges (e.g., 0 = "no correspondence"), which practice might be expected to increase interjudge reliability. The number of points on the rating scale ranged from two to eleven, with a mean of 4.2, and it is possible that the scales at the low end of the range may be too constrained to be sensitive, while those at the higher end require judges to make more fine judgements than is appropriate, and so may be insensitive in effect because they increase error variance. In this latter case of large rating range, interjudge reliability may be reduced. The same may be true of holistic rating scales, which ranged from 4 points to 101, and which were clearly reported as being labelled in only 14 out of the 52 studies in which a holistic scale was used. The number of items in the judging set may be a factor in determining the appropriate rating scale; in the studies surveyed, set size ranged from 2 to 36 items. Any future research which addresses the issue of the appropriate rating scale in this task could most usefully do so in the context of active training of judges, with feedback, in the use of such scales. Boerenkamp (1984) had considerable success in training eight independent judges to rate each statement made by a "psychic" about a missing person on a fully-labelled four-point scale of likelihood that it would apply to anyone in the population. To test the reliability of the judges' ratings, the judges were randomly assigned to two groups of four, and the average ratings of each statements were correlated, yielding correlations ranging from $r_s = +0.66$ (36df, $p<0.01$) to $r_s = +0.93$ (19df, $p<0.01$). The training consisted of having each judge rate independently the first statement in the transcript, followed by a discussion among the judges about the differences in their scoring. Then the second statement was scored and discussed, and so on, allowing the judges to discover why they differed from the group norm and to adjust their rating strategy accordingly. Similar training in rating statements for the likelihood that they were the product of deductive reasoning, also on a fully-labelled four-point scale, yielded similarly respectable interjudge reliabilities, ranging from $r_s = +0.66$ (72df, $p<0.01$) to $r_s = +0.95$ (20 df, $p<0.01$). Although no pretraining measures of reliability were taken, the assignment of ratings of the likelihood that a statement would be true of a person on a fully-labelled three-point scale by two untrained judges in a study by Tart and Smith (1968), showed perfect agreement only 49% of the time. The reliability of Boerenkamp's judges is relatively high compared to that generally obtained in free-response judging, and this may be a useful method for training the judges in the reliable assignment of ratings to mentation-target correspondence.

Maren (1986) discusses the application of artificial intelligence (AI) to give measures of the correspondence between target and mentation. However, she stresses that the development of appropriate AI systems is at an early stage. It seems that for the time being, the best bet for improving the reliability of atomistic (and possibly holistic) ratings may be the training of judges, with feedback, in the use of fully-labelled scales with a range appropriate for the task.

(ii) Weighting in accordance with the rarity of the mentation item's content

The probability of an exact match between a mentation item and an element of the target is equal to the product of the probability that the mentation item should occur on that trial out of all the others, and of the probability that that target element should be present in that target out of all other targets. This being so, the rarer the mentation item, the more weight it should receive. Stanford's response-bias hypothesis (1967), coming from a different angle, also suggests that rare responses should be relatively heavily weighted.

Although a number of experimenters have instructed judges to attach more weight to rare correspondences (e.g. Palmer, Khamashta, & Israelson, 1979; Sargent, 1980), the calculation of frequencies of mentation occurrence has been seldom. The exceptions are studies by Roll, 1971, Roll et al., 1973, and Tart and Smith (1968). In these studies, statements made by a medium about a number of people were weighted inversely according to the number of people in the study about whom the medium made the same statement.

Further work attempting to calculate norms for free-response mentation would need to take into account a number of factors. The setting may be important; in the ganzfeld, for example, the white noise often elicits imagery about waterfalls, beaches and aeroplanes. Some responses are common in certain states of consciousness, for example, faces are a common feature of hypnagogic imagery (Mavromatis, 1987). Presumably for this reason, Braud and Braud (1974) and Braud, Wood and Braud (1975) instructed their percipient (who later did the judging) to attempt to distinguish target-relevant impressions from those induced by the state itself (in this case, conventional meditation imagery).

As well as being dependent on the situation and state of consciousness of the percipients, mentation content frequencies may also vary from percipient to percipient; most experimenters will probably have come across percipients who in repeated testing, always mention one or more specific images which occur in each of their trials, while in contrast, Sargent, Bartlett and Moss (1982) reported that experienced participants in their ganzfeld study adopted the strategy of not bothering to report responses which they recognised as habitual. Frequency norms may also vary according to the nature of the target; percipients in a study in which geographical location is the target may be more inclined to talk about trees than a percipient in a study in which aspects of a person is the target.

(iii) Weighting in accordance with the rarity of the mentation item in target pool

As stated in (ii) above, the probability of an exact match between a mentation item and an element of the target decreases as the probability of the occurrence of the mentation item in the target pool decreases. Therefore, the rarer the mentation item in the target pool, the more heavily it should be weighted.

Jahn, Dunne and Jahn (1980) calculated the a priori probabilities of all values of their thirty binary descriptors in the entire target pool (i.e., the probability that people were present, the probability that people were absent, the probability that movement was present, etc.) to facilitate the heavier weighting of rarer target contents (or absence of content). In judging involving a judging set, frequency of content is usually neither calculated nor weighted.

(iv) Mentation characteristics

Several experimenters have attempted a formal analysis of which mentation categories, if any, tend to be psi-related. Sargent, Bartlett and Moss (1982), Sargent, Moss and Bartlett (1981), and Sargent, Milton, Payne and Bennet (1982) found that scoring on the basis of "bizarre" mentation was significantly above chance, although scoring on the basis of bizarre imagery was compared to the theoretical chance level, rather than to scoring on the basis of other imagery. Milton (1984) found significant psi-hitting on the basis of "surprising" imagery and Milton (1985) found significantly below-chance scoring on "fleeting" imagery according to the results of one of two independent judges. A third study by Milton (1987) examining a wide range of mentation categories found no significant results.

White, Krippner, Ullman and Honorton (1977) had one judge place mentation items from dream transcripts into one or more of seventeen categories, and had a second judge compare each item to the target for that night and mark it "telepathy present" or "telepathy absent". Eight categories (listed as waking imagery, hypnagogic and hypnopompic imagery, associations, colour, communication, witness, specificity and elaboration) were associated with target correspondences to a significant degree, the association with waking imagery being a negative one (i.e. waking imagery seemed to be associated with the target less often than chance). Seven categories (including anxiety, experiment-related, hostility-misfortune, penetration of self-boundaries, participation, vividness) yielded non-significant results, and there was insufficient data to test two categories (sex and violence). However, since it is not mentioned whether the mentation items were edited after being categorised, it may be that these results reflect at least in part the judges' expectations; he or she might have been more inclined to consider telepathy present for a mentation item which fell into a category which he or she expected to be successful.

In a study by Schouten and Merkestein (1985) percipients, already familiar with the target pool, had to record striking experiences during the day and later selected the day's target drawing from the pool on the basis of these experiences. In order to reduce the amount of work involved in the judging task, Merkestein selected for independent judging only those experiences which fell into five specified categories. Only memories yielded significant above-chance results ($p=0.000\ 08$), while spontaneous, unexpected inner experiences, dreams and daydreams, experiences related to a topic which the percipient had forgotten was in the target pool, and experiences related to the mood of the day, were not significantly target-related.

In addition to direct experimentation, some experimenters have offered anecdotal observations concerning what sort of mentation appears to be more psi-related than others. Reporting on an informal research discussion among apparently successful percipients and researchers, Schlitz (1984) notes that many of the participants felt that imagery which was fleeting, novel, or recurrent tended to be psi-related, and that nonvisual impressions, including kinaesthetic, auditory and olfactory images, were of equal or greater importance compared to visual imagery. Honorton and Harper (1974) observed that memory images seemed to be successful in a ganzfeld study; Dunne and Bisaha (1979) commented that logical inferences from an initial impression were unhelpful. In a ganzfeld experiment by Palmer, Bogart, Jones and Tart (1977), the scores of only one out of two independent judges yielded significant displacement scoring, and among the differences reported by the judges in their strategies was the inclination of the more "successful" judge

to pay more attention to "images that were unique in the general context of the subject's mentation, e.g., images that came as sudden breaks in the ongoing train of thought" (p.138).

Other experimenters have instructed judges to pay more attention to particular mentation categories. Thus Sargent (1980) instructed judges to pay more attention to mentations which the percipient reports as being novel, striking, odd, unusual, unexpected or particularly clear, and to pay less attention to mentations clearly linked with an immediate memory (thus not conflicting with the comment of Honorton and Harper (1974) above, which presumably relates mostly to long-term memories). The criteria for deciding whether or not one of Honorton's (1975) binary content categories is present in the mentation include "intensity" and persistence of content-related mentation.

It can be seen that a number of experimenters feel that certain mentation types may be more likely to be psi-related than others, although authors vary in their choices, and few seem to have based their opinions on formal research findings. This would seem to be another aspect of the judging process which would benefit from systematic, direct research, with anecdotal observations as a valuable starting point.

(v) Salient aspects of targets

In an ideal judging situation, those elements of the target material which are most salient for the percipient should be more heavily weighted in the judging than elements known not to be salient. For the purposes of this discussion, 'salient' describes an element of the target about which the percipient tends to give accurate information more often than chance. Thus if percipients were very often correct about whether or not people were present in a pictorial target, then mentation items dealing with the presence or absence of people should be more heavily weighted than other mentation items.

Roll et all (1973) applied such a weighting to mentation categories according to their content, made by a sensitive, and meant to apply to various people. The content categories were those of physical description, health, vocation/education, family, love life, future, wants, interests, needs, personal characteristics, and other, and mentation items of half the data were weighted in accordance with the success of mentation items in those content categories in the other half of the data.

The content categories used by Roll et al were presumably chosen since most of the sensitive's mentation could conveniently be coded in terms of them, rather than because each of these categories was believed to be highly salient; indeed, the study was partly one of salience. However, in studies where mentation and targets are coded in terms of content categories (e.g. Honorton, 1975; Jahn, Dunne and Jahn, 1980), content categories seem to be chosen not for their salience but for similar pragmatic reasons of allowing a fairly full description of the mentation report. Further research identifying salient content categories, to allow them to be weighted appropriately or used as the basis of coding systems, would be useful.

(vi) Correspondence types

A number of authors have discussed ways in which mentations have appeared to relate to targets in their studies, and some authors instruct their judges to watch out for some of these correspondences. Those mentioned have included literal, formal (shape), sensory (colour, material), symbolic/metaphorical, associational, emotional and functional correspondences, and it has been suggested that these correspondences may relate to either parts of or the

target, or to the whole, or both. However, authors differ, and sometimes conflict, in the importance they attach to these correspondence types. Some authors only take into account one or two types of correspondence, while others deal with most of them but weight heavily certain types which other authors feel are unimportant.

For example, Sargent, Bartlett and Moss (1982) in their judging instructions attach most importance to direct (presumably literal) correspondences, and then consider formal, associative, symbolic, and mood/emotive correspondences in order of decreasing importance (a similar order of importance is reflected in Sargent's (1982) judging instructions). These instructions conflict with the opinions of several other researchers, such as Dunne and Bisaha (1979) who consider that correspondences of shape, colour, size, and relation to other shapes, and metaphorical correspondences are more likely (and presumably more important) than literal correspondences. Similarly, Targ and Puthoff (1978) feel that correspondences of shape, form, colour and material are likely to be more accurate than correspondences of function or name; Schlitz and Haight (1984) instructed their judges to expect correspondences of shape or association rather than literal correspondences; Gelade and Harvie (1975) commented that accurate descriptions were rare, and that metaphorical and symbolic correspondences were more frequent; Hearne (1986) instructed independent judges to look particularly for symbolic correspondences, and Stanford (1979) used artists as judges on the basis of comments by other researchers indicating that meaning was often distorted in mentation but hat the form of the target was often described correctly. Other researchers, while instructing their judges about the types of possible correspondence have either urged their judges to give equal weight to all types, or have given instructions in which no type of correspondence was made to seem more important than any other (Moriarty and Murphy, 1967; Musso and Granero, 1973; Palmer, Khamashta and Israelson, 1979; Palmer, Bogart, Jones and Tart, 1977; and Wood, Kirk and Braud, 1977).

These differences among authors could be due to a number of different factors. Firstly, the type of correspondence thought to be most important in judging may relate to the percipient's mode of response, which tends to vary from study to study. Those experimenters who encourage their percipients to make drawings of their imagery have more opportunity to note correspondences (real or spurious) of form than meaning, while the reverse applies to those who encourage their percipients to make verbal responses. This may account for Sargent's preference of meaning over form in his ganzfeld studies in which percipients make mostly verbal responses, in contrast to the preference of form over meaning in the studies of, for example, Moriarty and Murphy (1967) and Musso and Granero (1973) which were both picture drawing studies.

A second factor in differences among authors may relate to individual differences between authors themselves, or between the percipients in those authors' studies. Hearne (1986) emphasised symbolic correspondences in his instructions to judges because the single percipient in that study seemed to have obtained such correspondences in earlier testing. Ullman (1966), discussing work on field dependence by Witkin (1965), suggested that the type of correspondence in each percipient's mentation might reflect whether the percipient is field dependent, with field dependent percipients yielding symbolic correspondences. In addition, in a study with both types of correspondence, the types of correspondence noted by the experimenter may depend in part on whether he or she is field dependent. The use of differed types of target material, may also result in different kinds of correspondence; for example, abstract art prints may yield correspondences of

form and sensory qualities, while pictorial representations of archetypes (such as those used by Gertz, 1983) may tend to yield symbolic correspondences.

PERCIPIENT JUDGES VERSUS INDEPENDENT JUDGES

So far, I have discussed the steps to be taken in an idealised judging process. A related issue is that of who is most likely to be suited to such a complex task. Most discussion in the literature on this issue has centred on the relative merits of percipients as judges of their own material, and of independent judges. The fact that at least one independent judge was used in 58.2% of the 98 studies in the database in which the use of an independent judge would have been appropriate may indicate a preference for independent judges.

Several reasons have been put forward for why independent judges should be preferred. First, the use of independent judges should give a uniformity of judging criteria across trials which may be lacking when percipients judge, resulting in reduced error variance with independent judges (Palmer, Bogart, Jones and Tart, 1977). Second, it should be easier to select or train a few good independent judges than to select numerous experimental participants who will be both good percipients and good judges (Palmer, Bogart, Jones and Tart, 1977). Third, the use of percipients as judges is likely to confound their ESP performance with their judging ability, such that relationships between their ESP score and other variables may be partly with their judging ability rather than their ESP; for example, a correlation between extraversion and the ESP measure may be due the extravert percipients judging more carefully to please the experimenter and hence increasing their ESP score (Stanford 1978, 1984). Fourth, the use of independent judges means that the percipient need only be shown the target at the end of the trial, which some experimenters feel may reduce the risk of precognitive displacement (Palmer, Bogart, Jones and Tart, 1977; Irwin, 1982). Fifth, independent judges are less likely to be ego-involved in the trial's outcome than the percipients since it is not their personal chance to demonstrate ESP in from of others, and may therefore be less likely to use such strategies as "going for broke" (i.e. artificially increasing the correspondence rating of a picture once they are sure it is the target, to make it look like a "better" hit) (Stanford and Sargent, 1983) or of deliberately avoiding giving points to a target which is a personal favorite (Sargent, 1980), although Stanford (1984) suggests avoiding telling independent judges that they are assessing ESP data in order to reduce the temptation for them also to "go for broke". Sixth, experienced independent judges may be more familiar with norms for free- response mentation and may be able to identify and hence weight appropriately mentation items which are unusual better than naive percipients. In a study by Sargent, Bartlett and Moss (1982), an independent judge who separated naive percipients' mentations into unusual and common mentations, obtained less of a scoring difference between the two than did the percipients who also categorised their own mentation. However, the judge also obtained lower scoring than the percipients in both categories, indicating that the judge may have been handicapped in the judging task (for example, by the percipients' inability to describe their imagery).

For reasons similar to those for avoiding percipient judging, several experimenters have explicitly recommended combining scores from several independent judges to dilute the effect of idiosyncrasies of each judge, such as the ability to only detect certain types of correspondence (Stanford, 1984)

or personal preferences for certain targets or mentation items which might influence the judge unduly (Targ and Targ, 1986). Some experimenters have judges working in consensus (e.g. Targ and Targ, 1986; Jahn, Dunne and Jahn, 1980), presumably for these reasons. Indeed, of those 57 studies in which at least one independent judge was used, only 20 used only one judge; the number used ranged from one to eight. However, the advantages of independent judges, multiple or otherwise, depends upon them being good judges, whether naturally, as a result of training, or due to the provision of full and appropriate instructions (Stanford, 1984).

The need to know what makes good judges and good judging has been stressed in the literature (Honorton and Stump, 1969; Sargent, 1980, 1981). Only two studies in those surveyed set out to compare the judging skills of judges of varying backgrounds. Roney-Dougal (1987) found that a psychotherapist independent judge with considerable experience in and knowledge of subliminal perception research scored most highly above chance (mean Z = +0.187, n.s.), while a "naive" poet scored slightly above chance (mean Z = +0.127, n.s.). A third judge who was a 'trained "psychic"' scored significantly below chance (mean Z = -0.16, t = 2.155, p = 0.04). This result is difficult to interpret, since we cannot know whether the percipients were "really" scoring above or below chance.

In a hypnotic dream study reported as a conference abstract, Keeling (1972) found that only a group of ten clinical psychology graduate students who judged the percipients' data obtained significantly above-chance scoring (p=0.018, one-tailed), while a group of ten undergraduates in an introductory psychology class, and a group of twenty middle-aged students in a YMCA course on the occult acting as judges yielded non-significant results. However, the results of the three groups were not strictly comparable, since the occult students judged different data from the other two groups, and the undergraduate psychology students did the judging in a different order from the other two groups.

Given the differences in scoring between the judges in these two studies, the judges' background and experience would seem to be an important variable in any free-response study. However, it was made clear in only 10 out of 57 studies using independent judges that the judge had experience relevant to judging (in areas of psychology, the visual or literary arts, etc. which deal specifically with the transformation of subconscious information, or previous experience of free-response judging).

No studies concerning the training of judges seem to have been made. However, the finding that Palmer, Bogart, Jones and Tart (1977) that a judge with experience of the ganzfeld gave significant evidence of displacement in a ganzfeld study while a second judge with no ganzfeld experience did not, might suggest that experience of the experimental procedure used in a study might usefully be included in any judge's training. Results from a study by Maher (1987), in which judges' scores increased with repeated judging of the same material presented in a different format each time, may suggest that simple repeated exposure to the judging task, or increasing familiarity with the judging material, may improve scoring.

The effect of instructions upon judges has similarly been a neglected topic, although Palmer, Khamashta and Israelson were led to compare the results of judging with and without instructions when they observed that uninstructed percipient judges scored below chance (Mean Z = -0.34), while two independent judges with full instructions scored above chance (mean Z = +0.37), the difference being significant (Wilcoxon T = 15, CR = -3.36, p<0.001, two-tailed). They had two more judges judge the data without

instructions, yielding a mean Z-score of +0.29, and concluded that the lack of judging instructions in this case probably did not cause the difference between the results of the percipients and the original two independent judges. However, the two uninstructed judges had taken part in a discussion of the judging of free-response material in Palmer's graduate class in parapsychology some months earlier, and so were not entirely naive. Instructions were reported as being given to judges without judging experience or knowledge of unconscious processes in only 12 out of 47 studies in which such judges were used. Further research is clearly needed on this topic.

The only reason against using independent judges has been that only the percipients themselves can have full knowledge of what their imagery really was and would be able to recognise personal symbolism (e.g. Palmer, 1986). This problem might also result in confounding the percipient's ability or inclination to fully report their imagery with their ESP performance if independent judging were used, possibly resulting in misleading relationships with other variables (Stanford, 1984). A number of experimenters have explored the importance of asking percipients for more information about their mentation after the end of the free-response period, by comparing the performance of independent judges provided with transcripts of the initial mentation reports, and with the initial transcripts plus additional information provided by the percipients.

Stanford (1984) has suggested training percipients in the reporting of imagery, while Palmer, Bogart, Jones and Tart (1977) suggest having an experimenter who is blind to the identity of the target, review the percipient's experience with him or her immediately after the response period and add to the transcript possibly relevant information (such as a full description of certain images, or the unusual qualities of images, phenomenological characteristics, and so on). Along these lines, it may also be advisable to offer percipients the opportunity to draw imagery which may have been difficult to describe verbally, or vice-versa, depending on the task.

Sondow (1979) found that independent judging by two experienced judges of the initial transcripts only from participants in a ganzfeld study yielded scoring exactly at chance (15 direct hits in 60 trials), while judging with the addition of the percipients' personal associations to the mentation gave significantly above-chance scoring (23 hits, $Z=2.39$, $p<0.02$). Each judge judged half of the initial transcripts only, and half of the transcripts with associations, so that no judge judged the same trial with and without associations. However, the percipients' judging yielded even higher scoring (30 hits in 60 trials).

In a dream study by Krippner, Honorton and Ullman (1972), independent judges judged first the initial mentation transcript alone, and then the transcript plus the results of an interview in which the percipient gave details of what mood accompanied the dream, what thoughts or memories it brought to mind, what elements of the dream made no sense in terms of the dreamer's personal life, and what the main them of the night's dreams had been. On the initial transcripts alone, the judges obtained two hits out of eight trials (with a one in eight chance of success). With the addition of the details of the interview, the judges obtained five hits, a result which was significantly above chance ($p=0.0012$, one-tailed). The percipient did not do any judging in this study, so no comparison with his scores can be made.

A similar procedure was used in a study by Ullman, Krippner and Feldstein (1966). In the interview, the percipient was asked what the dream reminded him or her of, what if anything seemed to be trying to intrude on the dream,

and whether there was anything in the dream which was different from the percipient's dreams, such as colour, feeling the dream to be real, or private symbolism. On the basis of the initial transcripts alone, the three judges (whose judging experience, if any, was not reported) scored significantly above chance ($F=8.30$, $p<0.01$); with the addition of the interview material, scoring was even higher ($F=18.14$, $p<0.001$). The percipient judging yielded non-significant results with the initial transcript alone, and results above chance at the $p<0.05$ level ($F=4.41$) with the addition of the interview material.

On the basis of these results, it seems that further elaboration by percipients on their initial mentation reports adds useful information, since scores with such elaboration were higher than those without in all three studies discussed above. However, while the percipients still managed to score at a higher level than the experienced judges in Sondow's study even when the judges were provided with their associations, Ullman, Krippner and Feldstein found that their (possibly inexperienced) independent judges scored higher than the percipient judges both with and without associations. This apparent conflict of findings may be in part due to the extra information which Ullman et al. elicited from their percipients during the interview. Clearly, more research needs to be directed to this question.

DISCUSSION

The most striking feature about judging practices in the literature surveyed is their variety, and in some aspects of judging, their contradiction. The level of description of aspects of judging is generally very brief, and it may be that judging practices are much more similar from laboratory to laboratory than appears in print. Similarly, the 4% of studies using independent judges which involved giving the judges full instructions concerning various types of transformation types along with detailed examples, may be an underestimate, since more experimenters may have given their judges equally full instructions without reporting it. However, either a lack of instructions or a lack of reporting them might imply a lack of importance being attached to the judging process within the field. Since judging is logically a crucial part of any free-response study, both more attention to judging and its reporting is surely merited. Delanoy (1987) has listed information about judging which should ideally be listed in any free-response study.

Although little direct research has been done on the judging process, the studies surveyed indicated many potentially profitable lines of research. The training of judges (real training with feedback, rather than merely repeated exposure to the judging task) has apparently not be explored, and may be a valuable research strategy in this area. Awareness of individual differences, methods of responding (verbal, pictorial, etc.), setting and target type are among the many variables which need to be considered in further research on judging, as well as aspects of procedure such as the use of rating scales with appropriate ranges and judging sets of an appropriate size for the task. We clearly need to know more about all aspects of judging as part of our efforts to improve the reliability and effectiveness of free-response experimentation in general.

REFERENCES

BOERENKAMP, H. G. (1984). Potential paranormal value of statements of psychics acquired under feedback conditions. European Journal of Parapsychology, 5, 101-124.

BRAUD, L. W., & BRAUD, W. G. (1974). Further studies of relaxation as a psi-conducive state. Journal of the American Society for Psychical Research, 68, 229-245.

BRAUD, W. G., & WOOD, R. (1977). The influence of immediate feedback on free-response GESP performance during Ganzfeld stimulation. Journal of the American Society for Psychical Research, 71, 409-427.

DELANOY, D. (1987). The reporting of methodology in ESP experiments. Parapsychology Review, 18, 1-4.

DUNNE, B. J., & BISAHA, J. P. (1979). Precognitive remote viewing in the Chicago area: A replication of the Stanford experiment. Journal of Parapsychology, 43, 17-30.

GELADE, G., & HARVIE, R. (1975). Confidence ratings in an ESP task using affective stimuli. Journal of the Society for Psychical Research, 48, 766, 209-219.

GERTZ, J. (1983). Hypnagogic fantasy, EEg, and psi performance in a single subject. Journal of the American Society for Psychical Research, 77, 155-170.

HEARNE, K. M. T. (1986). An analysis of premonitions deposited over one year, from an apparently gifted subject. Journal of the Society for Psychical Research, 53, 804, 376-382.

HONORTON, C. (1975). Objective determination of information rate in psi tasks with pictorial stimuli. Journal of the American Society for Psychical Research, 69, 353-359.

KEELING, K. R. (1972). Telepathic transmission in hypnotic dreams: An exploratory study. In Roll, W. G., Morris, R. L., & Morris, J. D. (Eds.), Proceedings of the Parapsychological no. 8, 1971.

HONORTON, C., & HARPER, S. (1974). Psi-mediated imagery and ideation in an experimental procedure for regulating perceptual input. Journal of the American Society for Psychical Research, 68, 156-168.

HONORTON, C., & STUMP, J. P. (1969). A preliminary study of hypnotically-induced clairvoyant dreams. Journal of the American Society for Psychical Research, 63, 175-184.

IRWIN, C. P. (1982). The role of memory in free-response ESP studies: Is target familiarity reflected in the scores? Journal of the American Society for Psychical Research, 76, 1-22.

JAHN, R. G., DUNNE, B. J., & JAHN, E. G. (1980). Analytical judging procedure for remote perception experiments. Journal of Parapsychology, 44, 207-231.

KEELING, K. R. (1972). Telepathic transmission in hypnotic dreams: An exploratory study. In Roll, W. G., Morris, R. L., & Morris, J. D. (Eds.), Proceedings of the Parapsychological no. 8, 1971.

KRIPPNER, S., HONORTON, C., & ULLMAN, M. (1972). A second precognitive dream study with Malcolm Bessent. Journal of the American Society for Psychical Research, 66, 269-279.

MAHER, M. (1987). Replication of an "incline" effect in blind judging scores. In Weiner, D. H., & Nelson, R. D. (Eds.), Research in Parapsychology 1986, Scarecrow Press: Metuchen, N.J.

MAREN, A. J. (1987). Representation and performance evaluation approaches in psi free-response tasks. In Weiner, D. H., & Nelson, R. D. (Eds.), Research in Parapsychology 1986, Scarecrow Press: Metuchen, N.J.

MAVROMATIS, A. (1987). Hypnagogia. Routledge & Kegan Paul: London & New York.

MILTON, J. (1984). The effect of the presence of an agent on ESP performance and of the isolation of the target from its controls on displacement in a ganzfeld clairvoyance experiment. In White, R. A., & Broughton, R. S. (Eds.), Research in Parapsychology 1983, Scarecrow Press: Metuchen, N.J.

MILTON, J. (1985). The effect of agent strategies on the percipient's experience in the ganzfeld. In White, R. A., & Solfvin, J. (Eds.), Research in Parapsychology 1984, Scarecrow Press: Metuchen, N.J.

MILTON, J. (1987). Judging strategies to improve scoring in the ganzfeld. In Weiner, D. H., & Nelson, R. D. (Eds.), Research in Parapsychology 1986, Scarecrow Press: Metuchen, N.J.

MORIARTY, A. E., & MURPHY, G. (1967). An experimental study of ESP potential and its relationship to creativity in a group of normal children. Journal of the American Society for Psychical Research, 61, 326-338.

MUSSO, J. R., & GRANERO, M. (1973). An ESP drawing experiment with a high-scoring subject. Journal of Parapsychology, 37, 13-36.

PALMER, J. (1986). Experimental methods in ESP research. Chapter in Edge, H., Morris, R. L., Palmer, J., & Rush, J. H., Foundations of parapsychology, (pp. 111-137), Routledge & Kegan Paul: Boston, London & Henley.

PALMER, J., BOGART, D. N., JONES, S. M., & TART, C. T. (1977). Scoring patterns in an ESP ganzfeld experiment. Journal of the American Society for Psychical Research, 71, 121-145.

PALMER, J., KHAMASHTA, K., & ISRAELSON, K. (1979). An ESP ganzfeld experiment with transcendental meditators. Journal of the American Society for Psychical Research, 73, 333-348.

ROLL, W. G. (1971). Free verbal response and identi-kit tests with a medium. Journal of the American Society for Psychical Research, 65, 185-191.

ROLL, W.G., MORRIS, R. L., DAMGAARD, J. A., KLEIN, J., & ROLL, M. (1973). Free verbal response experiments with Lalsingh Harribance. Journal of the American Society for Psychical Research, 67, 197-207.

RONEY-DOUGAL, S. M. (1987). A comparison of subliminal and psi perception: Exploratory and follow-up studies. Journal of the American Society for Psychical Research, 81, 141-181.

SARGENT, C. L. (1980). Exploring psi in the ganzfeld. Parapsychology Foundation: New York, N.Y.

SARGENT, C. L. (1981). ESP in the twilight zone: State of the art. Parapsychology Review, 12, 1-7.

SARGENT, C. L. (1982). A ganzfeld GESP experiment with visiting subjects. Journal of the Society for Psychical Research, 51, 790, 222-232.

SARGENT, C. L., BARTLETT, H. J., & MOSS, S. P. (1982). Response structure and temporal incline in ganzfeld free-response GESP testing. Journal of Parapsychology, 46, 85-110.

SARGENT, C. L., MILTON, J., PAYNE, J., & BENNET, S. (1982). Unpublished study.

SARGENT, C. L., MOSS, S. P., & BARTLETT, H. J. (1982). Unpublished study.

SCHLITZ, M. (1984). Esalen meetings on psi research. Parapsychology Review, 15, 10-12.

SCHLITZ, M. J., & HAIGHT, J. (1984). Remote viewing revisited: An intrasubject replication. Journal of Parapsychology, 48, 39-49.

SCHOUTEN, S. A., & MERKESTEIN, J. (1985). A free-response study in a real-life setting. European Journal of Parapsychology, 6, 19-32.

SONDOW, N. (1979). Effects of associations and feedback on psi in the ganzfeld: Is there more than meets the judge's eye? Journal of the American Society for Psychical Research, 73, 123-150.

STANFORD, R. G. (1967). Response bias and the correctness of ESP test responses. Journal of Parapsychology, 31, 280-289.

STANFORD, R. G. (1978). Special problem areas in research methodology. In W. G. Roll (Ed.), Research in Parapsychology 1977, Scarecrow Press: Metuchen, N. J.

STANFORD, R. G. (1984). Recent ganzfeld-ESP research: A survey and critical analysis. In Krippner, S. (Ed.), Advances in Parapsychological Research 4. McFarland: Jefferson, N. C.

STANFORD, R. G. (1979). The influence of auditory ganzfeld characteristics upon free-response ESP performance. Journal of the American Society for Psychical Research, 73, 253-272.

STANFORD, R. G., & SARGENT, C. L. (1983). Z scores in free-response methodology: Comments on their utility and correction of an error. Journal of the American Society for Psychical Research, 77, 319-326.

TARG, R., & PUTHOFF, H. E. (1977). Mind-Reach. Dell: New York.

TARG, E., & TARG, R. (1986). Accuracy of paranormal perception as a function of varying target probabilities. Journal of Parapsychology, 50, 17-27.

TART, C. T., & SMITH, J. (1968). Two token object studies with Peter Hurkos. Journal of the American Society for Psychical Research, 62, 143-157.

ULLMAN, M. (1966). A nocturnal approach to psi. In Roll, W. G. (Ed.), Proceedings of the Parapsychological Association, no. 3, 1966.

ULLMAN, M., KRIPPNER, S., & FELDSTEIN, S. (1966). Experimentally-induced telepathic dreams: Two studies using EEG-REM monitoring techniques. International Journal of Parapsychology, 8, 577-603.

WHITE, R. A., KRIPPNER, S., ULLMAN, M., & HONORTON, C. (1971). Experimentally-induced telepathic dreams with EEG-REM monitoring: Some manifest content variables related to psi operation. In Roll, W. G., Morris, R. L., & Morris, J. D. (Eds.), Proceedings of the Parapsychological Association no. 5, 1968.

WOOD, R., KIRK, J., & BRAUD, W. (1977). Free response GESP performance following ganzfeld stimulation versus induced relaxation, with verbalised versus nonverbalised mentation: A failure to replicate. European Journal of Parapsychology, 1, 80-93.

WITKIN, H. A. (1965). Psychological differentiation and forms of pathology. Journal of Abnormal Psychology, 10, 317-336.