President's letter: Small Wonder

The outcomes of parapsychology experiments are often portrayed as being very small in comparison with other areas of research, perhaps achieving 'statistical significance' but being too weak to be of any real-world relevance. Rebus and Alcock[i] are fairly typical in claiming that "because experiments produced only null or marginally significant results with small effect sizes, parapsychologists [have] resorted to meta-analyses where the results from large numbers of studies are combined on the presumption that psi effects will be uncovered with this increased statistical power." Similarly, French observes,[ii] "It is sometimes argued by critics that the effect sizes typically found in parapsychology are so tiny that no sensible person would choose to consider them paranormal as opposed to being due to some (possibly unknown) combination of ordinary factors". Even some researchers who are sympathetic to the objectives of parapsychology are willing to concede this point. For example, Kennedy asserts,[iii] "one of the most important and perplexing questions in parapsychology is why psi phenomena are so weak", while Goertzel and Goertzel admit,[iv] "It seems clear from the available data that, if psi is a real phenomenon, it is generally weak and finicky, and varies based on a host of hard-to-pin-down variables… Arguably, the most important challenge for psi researchers is to increase the effect size in their experiments". In this Note I'll take a closer look at the effect sizes that are reported in parapsychological research to see whether they really are too small to be indicative of real effects. In particular, I will compare them with findings from other areas of psychology to see whether they are deviant in the way that commentators have suggested (a fair comparison I would argue, given that both involve capturing the 'performance' of human participants).

First it would be useful to have a sense of what effect sizes are and how they are interpreted. Experimentation is essentially a quantitative exercise – the outcomes are converted into numbers that might reflect absolute values (such as number of words recalled correctly, or reaction time in milliseconds to respond to a stimulus) or relative values (such as participant ratings of how pleasant a task was, or what extraversion score they would give to someone described in a short vignette). These values are combined across large numbers of participants in statistical analysis that allow us to see whether scores are collectively higher in one situation than in another (e.g., do people recall more words with one memory technique compared with another), or to see if performance is better than we would expect just by chance (e.g., do people presented with words so briefly that they report only seeing a flash of light nevertheless pick out the presented word more often than chance would allow when asked to 'guess' which of a set of words was presented).

There is a wide range of statistical tests that can be used to make those comparisons, depending on the measurements taken, on the conditions of the experiment, but also on researcher preference, so that it can be difficult to tell whether different studies have found similar effects. Meta-analyses that combine different studies together will convert all the various statistical outcomes into a standard effect size, such as Pearson's r or Cohen's d. These typically compute the size of the difference between groups (M_1 and M_2 in Figure 1) in relation to how much scores vary within a group due to other factors (SD_1 and SD_2) — the larger the difference, the greater the effect. McLeod has provided a useful guide (Table 1)[v] for how various sizes of effect can be understood in terms of the proportion of people in the experimental group who outperform those in the control group. This table uses Cohen's original description of effects as 'small', 'medium' and 'large'. A medium effect is "likely to be visible to the naked eye of a careful observer" and other effects are described in relation to it, so a

small effect is unlikely to be obvious to a careful observer but still is not so small as to be trivial. A large effect may be so obvious as to not need statistical analysis to confirm it. Jessica Utts helpfully illustrates this with the example of heights. The standard deviation for adult height is about 2.5 inches for both men and women. If the average difference in height between men and women gave only a small effect size of 0.2, then that would equal about half an inch (0.2 x 2.5). In practice it would be very difficult to detect such a small height difference even if we were to observe lots of men and women. In the UK the actual average height of men is 5ft 9in and of women is 5ft 3in, giving an effect size of 2.4, which is so large that we should immediately notice the sex difference in height.

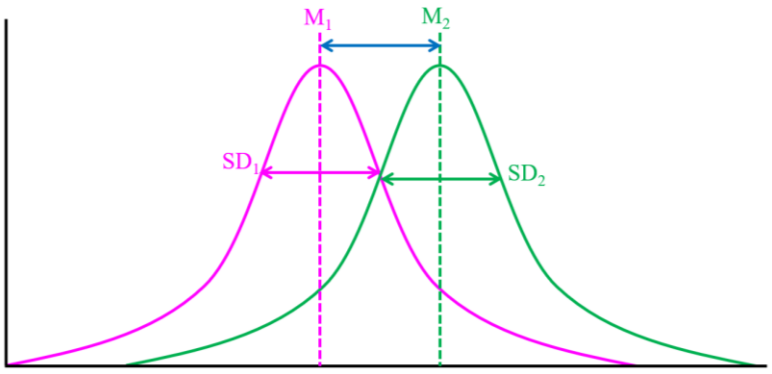

Figure 1: Distribution of scores for a control group and an experimental group

Relative size	Effect size, d	% of the control group who fall below the mean of the experimental group
	0.0	50%
Small	0.2	58%
Medium	0.5	69%
Large	0.8	79%
	1.4	92%

Table 1: converting effect sizes into group differences

So, where do parapsychological effects fit in this scheme? In Table 2 I have reproduced some of the summary findings reported by Cardeña (discussed in more detail in previous President's Notes). It is clear that most effects fall in the range of about .1 to .2, so would be classed as small or very small. Some effects (Forced choice ESP, PK Dice) seem so tiny as to be negligible when it comes to affecting real-world situations. At this point, then, concerns about parapsychology producing only modest effects seems justified. However, the picture changes when we look at other effect sizes in psychology. The Open Science Collaboration, a coalition of 270 research psychologists, attempted to replicate findings from 100 studies reported in high-ranked psychology journals, but only 47.4% of the results were close enough to the original findings to be interpreted as successful. The mean effect size of the 100 original studies was r = 0.40[vi] but this fell to 0.20 in the replication studies. Maybe these elite publications are not representative of psychology as a whole. Schäfer T and Schwarz[vii] analysed a random selection of 100 published empirical studies from each of 9 domains of psychology (biological, clinical, developmental, etc.), and found that the median effect size, r, was .36; however, the value for

studies that were pre-registered (and so were less susceptible to publication bias or selective reporting) gave a median effect of only 0.16. The largest effects were in disciplines that benefited from more systematic methods and instrumentation, such as biological psychology, whereas disciplines such as social and developmental psychology produced much smaller effects. In this context, then, it looks as if the reported effects in parapsychology are broadly on a par with many other subdisciplines of psychology. Claims to real-world relevance seem reasonable insofar as these other areas also lay a claim to it.

Type of experiment	No. of experiments	Effect size	significance
Ganzfeld	108	.14	$< .10^{-16}$
Precognition / Bem-type	90	.09	1.2×10^{-10}
Psi dream	52	.18	2.7×10^{-7}
Remote viewing (Dunne & Jahn)	88	.21	3×10^{-8}
Presentiment	26	.21	5.7×10^{-8}
Forced choice ESP	309	.02	6.3×10^{-25}
DMILS	36	.11	<.001
Remote staring	15	.13	.001
Attention facilitation	11	.11	.029
PK Dice	73	.01	<.001

Table 2: Summary effect sizes for the main lines of experimental parapsychology

Finally, I would like to echo the argument that even relatively small effects may still have important implications. Funder and Ozer[viii] describe an effect-size r of .05 as an "effect that may be very small for the explanation of single events but potentially consequential in the not-very-long run". This is an important distinction; we may not be able to make confident predictions about how psi might affect a particular individual or be evident in a particular situation, but we might be confident that it can have far reaching consequences for the population as a whole across the whole range of situations they may find themselves in. To illustrate this point, Jessica Utts[ix] has often cited a medical study that tested whether 325 mg of aspirin taken every other day could reduce mortality from cardiovascular disease. Participants were 22,071 apparently healthy US male physicians aged 40–84 years at entry, so this was a very large study. After five years of treatment and follow-up, it was found that there were 17.13 heart attacks per 1,000 in the group taking the placebo but only 9.42 heart attacks per 1,000 in the group taking aspirin. The study was terminated early because the oversight committee felt it was unethical to withhold the benefits of treatment from the control group. Rosenthal and Rosnow[x] have calculated the effect size r as .034 (which equates to d = .068)[xi], much smaller than the effects found in experimental parapsychology. This is not an isolated case; Rosenthal and Rosnow list a range of intervention effects for conditions like polio, convulsions, blood clots and AIDS that produce similar effect sizes. "One result of our consideration of these biomedical effect size estimates," they conclude, "is to make us more sanguine about the magnitude and importance of research findings in the behavioural and social sciences."

---

[i] A.S. Reber, & J. E. Alcock (2020). Searching for the impossible: Parapsychology's elusive quest. *American Psychologist*, *75*(3), 391–399. https://doi.org/10.1037/amp0000486

[ii] C. French (2010). Reflections of a (relatively) moderate skeptic. In S. Krippner & H. L. Friedman (Eds.) *Debating psychic experience*. Praeger. (pp. 53-64)

[iii] J. E. Kennedy, (2001). Why is psi so elusive? A review and proposed model. *Journal of Parapsychology*, *65*, 219–246.

[iv] T. Goertzel & B. Goertzel (2015). Skeptical responses to psi research. In Goertzel and Goertzel (Eds.), *Evidence for psi: Thirteen empirical research reports*. McFarland & Co. (pp. 291-301).

[v] S. McLeod (2019). What does effect size tell you? *Simply Psychology*: https://www.simplypsychology.org/effect-size.html

[vi] As a rough rule of thumb r values are about half the size of comparable d values, with an r of 0.1 equalling a small effect size, 0.3 a medium one, and 0.5 a large one.

[vii] T. Schäfer & M.A. Schwarz (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, 10, 813. doi:10.3389/fpsyg.2019.00813

[viii] D. Funder & D.J. Ozer (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168.

[ix] J. Utts (1999). The Significance of Statistics in Mind-Matter Research. *Journal of Scientific Exploration*, *13*(4), 615-638.

[x] R. Rosnow & R. Rosenthal (2003). Effect Sizes for Experimenting Psychologists. *Canadian Journal of Experimental Psychology*, 57(3), 221-237.

[xi] M. Borenstein, L.V. Hedges, J.P.T. Higgins, & H.R. Rothstein (2009). *Introduction to Meta-Analysis*. Chapter 7: Converting Among Effect Sizes. John Wiley & Sons, Ltd