Could the suppression of unsuccessful pilot studies explain the "Feeling the Future" results?

Daryl Bem's paper, "Feeling the Future" (2011), presented ten sets of data from nine separate experiments which appeared to show statistically strong evidence for retroactive influence - that is, the influence of unpredictable future events on past ones. These experiments were based on standard protocols from experimental psychology, but with the sequence of events altered in order to test for time-reversed versions of established effects.

These findings were controversial, and critics proposed a number of ways in which they could be explained in conventional terms. One of these suggestions was made by Ulrich Schimmack (2018a) after Bem provided him with copies of the experimental data, which he reanalysed. Schimmack discovered a decline effect, in which earlier participants within each experiment tended to perform more strongly than later ones. He suggested that this might reflect a situation in which a number of shorter pilot studies had been performed, and only the more successful ones had been continued as full experiments, with the data from unsuccessful pilot studies being discarded.

The size of the decline effect

The nine experiments presented by Bem contained ten separate data sets, and each had between 50 and 200 participants, who each contributed one session of experimental trials. The main results of Schimmack's analysis were presented in the form of a table showing separate p-values[1] for the first 50 and second 50 sessions in each data set.[2]

The statistical analysis of these experiments is not entirely straightforward. In Bem's paper the main results were calculated using *t* tests, partly with a view to encouraging replication by keeping the analysis as simple and familiar as possible.[3] Strictly, this procedure is justified only when the measurements being analysed are normally distributed. This condition would be satisfied only approximately in these experiments. In five of them, the participants' scores were based on the number of successes in a series of binary trials. In two others (Experiments 8 and 9), the scores were based on the numbers of words recalled in two different categories. In two more (Experiments 3 and 4), it was based on the difference of average response times for two conditions.

So only in two experiments out of nine was the score a continuous variable, and even then it would not necessarily have been normally distributed. For the majority of the experiments - the ones consisting of binary trials - the results can be analysed exactly (in the absence of a psi effect) using the binomial distribution. In these cases, Bem also presented the exact binomial *p*-values, which showed that the use of the *t* test resulted in only a relatively small error. But this error would be expected to be larger if only smaller subsets of the participants were considered. For the seven experiments where the scores were discretely rather than continuously distributed, there is likely to be a tendency for the *t* test to overestimate the statistical significance of departures from expectation.

Schimmack followed Bem in using *t* tests to calculate his *p*-values. One-tailed tests were used, because in the hypotheses as stated by Bem all the effects were expected to work in a particular direction. The table below shows the *p*-values presented by Schimmack, and a second set of *p*-values that have been corrected in three respects: (1) for the three data series of experiments 5 and

[1] The p-value is the probability that the result observed, or a result that departed even more strongly from expectation, would have happened by chance in the absence of a psi effect. A p-value is said to be statistically significant if it falls below some specified level, most commonly 0.05.

[2] Note that Experiment 9 had only 50 participants, while Experiments 3 and 4 had 97 and 99 respectively.

[3] Bem (2011), pages 407 and 408.

6, Schimmack calculated the values from the difference between experimental and control trials, rather than between experimental trials and chance expectation (see Schimmack, 2018b), (2) in the data file for Experiment 7 the sessions were out of date order, so that the first 50 entries were not the earliest chronologically, and (3) in four other cases Schimmack appears to have used a two-tailed rather than a one-tailed test, or considered the wrong tail[4]. The table also gives exact *p*-values calculated using the binomial distribution for the binary trials, and a discrete distribution for the two word-recall experiments, which will still not be exact but should be more accurate than the *t* test.[5]

Table. *p*-values calculated for the first 50 and second 50 sessions of each data series in Bem's experiments. Calculated using either the *t* test (with Schimmack's original figures in parenthesis) or discrete probability distributions where available.

	t Test		**Discrete probability distributions where available**	
Experiment	**Sessions 1-50**	**Sessions 51-100**	**Sessions 1-50**	**Sessions 51-100**
1	0.004	0.194	0.007	0.184
2	0.096	0.170	0.097	0.192
3	0.020 (0.039)	0.100	0.020	0.100
4	0.033	0.067	0.033	0.067
5	0.010 (0.013)	0.133 (0.069)	0.020	0.145
6 (negative trials)	0.171 (0.412)	0.242 (0.126)	0.188	0.224
6 (erotic trials)	0.033 (0.023)	0.276 (0.410)	0.056	0.345
7	0.039 (0.020)	0.599 (0.338)	0.056	0.602
8	0.010	0.318	0.010	0.296
9	0.003	NA	0.015	NA

The corrections to Schimmack's *p*-values make no difference to which of them are statistically significant (on a criterion of $p<=0.05$). But the alternative *p*-values based on discrete distributions

[4] A two-tailed test appears to have been used for Experiment 3, sessions 1-50, and Experiment 6 (erotic trials), sessions 1-50. The wrong tail seems to have been considered for Experiment 6 (negative trials), sessions 1-50, and Experiment 7, sessions 51-100. This may have been the result of halving a two-tailed *p*-value, as the effect was in the direction opposite to that expected in these two cases, so that this would give the one-tailed *p*-value for the opposite tail.

[5] The discrete distribution gives the probability of obtaining the observed score or a better one, given the total number of words recalled, calculated by exact computation of the number of ways of choosing the practice set of words. All the possible choices for the practice set were included. Therefore this distribution is not exactly applicable to the experiment, because the words were subdivided into four categories, and the practice set was constrained to contain half the words within each category.

do make some difference. Where Schimmack's *p*-values are significant, the alternative values tend to be larger, reflecting a tendency of the *t* test to overestimate significance.[6] In two cases where Schimmack's *p*-values are significant - for participants 1-50 in Experiments 6 (erotic trials) and 7 - the exact *p*-values based on the binomial distribution are not. However, even using the discrete distributions where available, the contrast between early and late participants remains noticeable - of the nine data sets that extend beyond 50 participants, five are significant for the first 50 participants, but none are significant for the second 50.

The difference between *t* tests and discrete distributions is more pronounced for smaller subsets of participants. Schimmack remarks that nine of the ten data sets reached significance at or before the fifteenth participant. Athough this is true when *t* tests are used, if instead discrete distributions are used where available, only five out of ten reach significance by that point.[7]

Nevertheless, there seems very little doubt that there is a real decline effect in the data. If the nine data sets that continued beyond 50 sessions are considered, and the alternative discrete statistics are used where available, it is possible to make an overall comparison between the first 50 sessions and the second 50. This can be done by converting each *p*-value to a *z* statistic (the corresponding value of a standard normal variable), averaging *z* over all the data series for each group of sessions, and then finding the difference between the values of *z* for the two groups. Using a normal approximation, the decline between early and late sessions is associated with a (one-tailed) *p*-value of 0.017 (which would remain significant for a two-tailed test for any difference between the groups, rather than specifically a decline).

Could the suppression of unsuccessful pilot studies explain the results?

Decline effects of one kind or another are well known in parapsychology. They have often been seen by sceptics as a sign that successful studies are the result of "questionable research practices" rather than genuine psi effects. In contrast, parapsychologists have pointed to other potential reasons for decline - either conventional ones of a kind that may be seen in any psychological experiment, or else something intrinsic to the nature of psi. One frequent sceptical claim is that decline effects are often seen after new research paradigms are introduced. The suggestion is that success may happen by chance in initial exploratory studies, but that afterwards these cannot be replicated, and there is a reversion to the mean.

This is essentially Schimmack's suggestion regarding Bem's experiments. He proposes that if a number of pilot studies had been conducted, some of them would have produced strong results by chance. If these were continued to completion and the others were discarded, the result would be that in each experiment early participants performed better than later ones. However, Schimmack also observes that the later participants considered in his analysis (the second 50 in each experiment) still performed better than would be expected by chance.[8] He says this suggests that "even these trials are biased by selection for significance". So apparently he envisages a second stage of selection, where even some experiments that had been continued after the pilot stage were

[6] Out of eight cases in which Schimmack's *p*-value was significant, the alternative *p*-value was smaller in only one, approximately the same in two, and larger in five.

[7] These five include the two priming experiments, for which discrete distributions are not available. But despite the fact that, strictly speaking, the *t* test will not be valid, the finding of significance does appear secure. If instead a non-parametric test is used (the Wilcoxon signed-rank test, in which a sum of the ranks of the differences is calculated, with the signs determined by the direction of the difference), both these data sets still attain significance. This is despite the fact that the Wilcoxon statistic is a discrete one, and will tend to underestimate significance when applied to continuous measurements.

[8] Schimmack's analysis covered only the first 100 sessions in each experiment. But the same was true of the results beyond this point, in the four data series that contain more than 100 sessions (all of which were based on binary trials). For them, the success rate beyond the 100th session was 51.6%, which is associated with a *p*-value of 0.007.

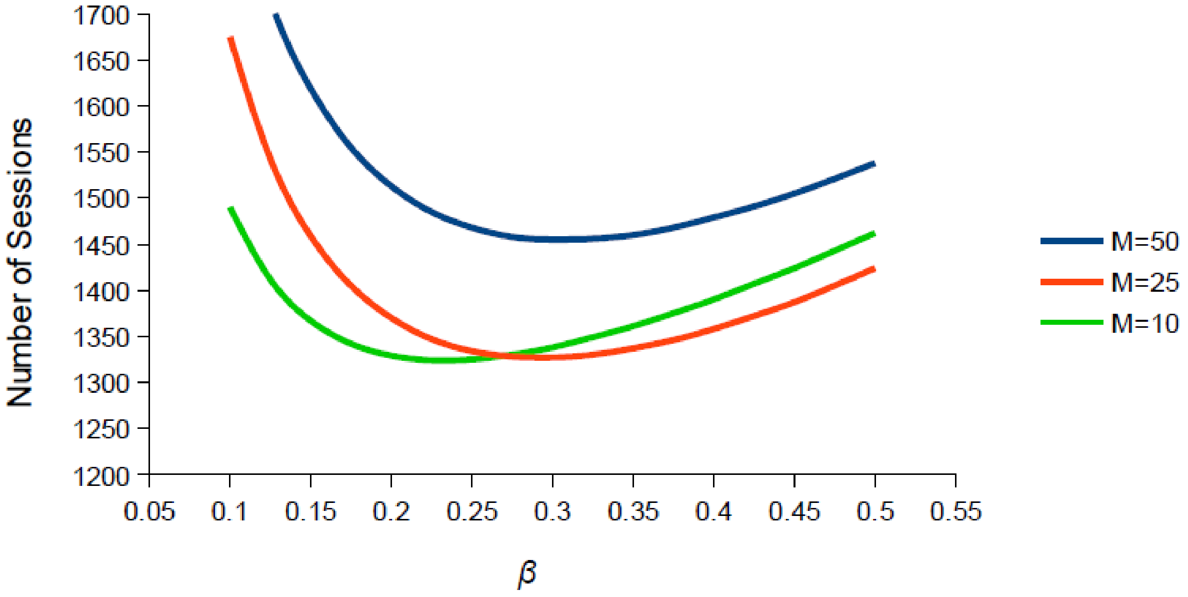
discarded because the results were not strong enough.

Could such a scenario plausibly account for the results published by Bem? It should be made clear that in subsequent email correspondence with Schimmack, Bem denied that anything like that had happened. He insisted that he had maintained a distinction between pilot studies performed on himself and his graduate students, and formal studies on real participants, and he denied that he had discarded failed experiments (Schimmack, 2018b). But apart from that, how believable is the suggestion on its own terms?

One obvious question is just how many data would have had to be generated - and how many suppressed - in order to produce the published results. A rough answer to that question can be found by using a very simple model of Schimmack's scenario, as follows. Suppose that a complete experiment consists of 100 sessions, that a pilot experiment consists of only *M* sessions (for some fixed number *M*), and that the score for each session can be approximated by a normally distributed random variable. For each experiment, once the *M* sessions have been done, the results are examined and the data are discarded if they fail to attain statistical significance at some specified level β. Otherwise the experiment is continued to completion, but the completed experiment is also discarded if it fails to achieve significance at some other specified level α. Using this model we can calculate numerically how many individual sessions will be required, on average, to produce one successful experiment.

For example, if the required significance for an experiment of 100 sessions is set at α=0.05, the plot below shows the average number of sessions that would be required for one successful experiment, for three choices of the pilot study size *M* (10, 25 and 50 sessions) and a range of values of the required significance β for each pilot study.[9]

Figure. Average number of sessions required for one successful experiment of 100 sessions, as a function of β, with α=0.05, for three values of *M* (10, 25 and 50).



[9] If the hypothetical selection of pilot studies were based on *t* tests, there might be a tendency to overestimate their significance somewhat. But, as a range of values of β is being considered here, that would not affect the implications of this model.

Further numerical investigation shows that there exist optimal choices for M and β - namely, the choices that will reduce to a minimum the average number of sessions required to produce one successful experiment. When α=0.05, the optimal choices are M=16 and β=0.266, and on average about 1,311 sessions would be required for a successful experiment. Of these sessions, only 100 would be retained, so 1,211 - or about 92% of all the experimental data - would have to be discarded.

If the alternative discrete p-values are used where available, five of Bem's experiments are found to be significant at the α=0.05 level by the 100th session.[10] But it is worth noting that all five of these are also found to be significant at the more stringent level of α=0.02. A calculation for this value of α shows that the optimal choices are M=16 and β=0.220, and that on average about 2,954 sessions would be required for a successful experiment. Of these, 2,854 - or about 97% - would have to be be discarded.

These figures can be compared to those that would apply for the more straightforward strategy of simply performing a large number of experiments and retaining only those with significant results. For α=0.05, that would involve discarding 95% of sessions, and for α=0.02, discarding or 98%. Although slightly fewer sessions are required in the scenario involving pilot studies - so that, in the optimal situation, only 92% and 97% of the sessions would have to be discarded - the reduction is relatively small. Given that anyone discarding such a large proportion of experimental data could be under no illusion about the impropriety of their actions, it is difficult to understand why the complicated and laborious scheme suggested by Schimmack should have been adopted.

There is one other surprising fact that is difficult to reconcile with the idea that the decline effect in Bem's results is the result of data selection.

In the two priming experiments, each participant also performed trials to test for the conventional (forward) priming effect, as well as the time-reversed (retroactive) one. In experiment 3, comparison of the results for sessions 1-50 with those for session 51-97 shows a decline effect for forward priming as well as for retroactive priming. In fact, the decline for forward priming was the steepest seen in the whole experimental series - in the earlier sessions it was by far the strongest of all the effects seen, but for the later ones it was not even statistically significant. By the same method used above to characterise the overall decline effect for the retroactive data series, the (one-tailed) p-value associated with this decline was 0.034. This was the only one of all the data series that showed a significant decline when analysed individually.

Obviously it is very hard to understand how the hypothetical selection of data favourable to a retroactive priming effect could produce the spurious appearance of a decline in a real forward priming effect in the same participants. This suggests that the decline observed for the conventional (forward) priming process is a real one, and that it may be relevant to the declines seen in the retroactive processes.

Chris Phillips, 9 April 2021

[10] Recalling that Experiments 3 and 4 contained only 97 and 99 sessions respectively.

References

Daryl J. Bem (2011). Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect. *Journal of Personality and Social Psychology*, 100(3), 407-425.

Ulrich Schimmack (2018a). Why the Journal of Personality and Social Psychology Should Retract Article DOI: 10.1037/a0021524 "Feeling the Future: Experimental evidence for anomalous retroactive influences on cognition and affect" by Daryl J. Bem. Blog post at https://replicationindex.com/2018/01/05/bem-retraction/

Ulrich Schimmack (2018b). My email correspondence with Daryl J. Bem about the data for his 2011 article "Feeling the future". Blog post at https://replicationindex.com/2018/01/20/my-email-correspondence-with-daryl-j-bem-about-the-data-for-his-2011-article-feeling-the-future/