Journal of Parapsychology, 80(2), 156–168

IS THE METHODOLOGICAL REVOLUTION IN PSYCHOLOGY OVER OR JUST BEGINNING?

By J. E. Kennedy

ABSTRACT: Significant results from parapsychological experiments using standard psychological research methods motivated psychologists to recognize some widespread methodological deficiencies and the need for preregistered well-powered confirmatory research. Psychological researchers have not yet recognized several other common methodological weaknesses that can be expected to cause this cycle to be repeated. When confronted with the choice between psi versus overlooked methodological deficiencies, psychologists will recognize the need for methodological improvements. These overlooked methodological factors include: (a) deficient study registration practices, (b) bias from dropouts and incomplete data, (c) the need for software validation, (d) measures to prevent experimenter fraud, (e) appropriate statistical methods for confirmatory research, (f) failure to consider inferential errors with Bayesian analyses, (g) the weaknesses of retrospective meta-analysis and strengths of prospective meta-analysis, and (h) problems from statistical dependence for the outcome variables in statistical analyses. Psychological and parapsychological researchers can easily avoid this inefficient process of methodological evolution driven by controversies about parapsychological findings. Research practices that address these methodological deficiencies are available and will eventually be recognized as needed for psychological and parapsychological research. Recommended practices for addressing these methodological weaknesses are described.

Keywords: research methodology, incomplete data, software validation, experimenter fraud, statistical power

Parapsychologists have often argued that their research methods have been as good as the standard methods used in psychological research (Cardeña, Palmer, & Marcusson-Clavertz, 2015, p. 1). With the publication of Bem's (2011) paper on "feeling the future," psychological researchers finally paid attention to this point and agreed. However, their conclusion was not that the research provided evidence for psi, but that the methods for psychological research needed to be substantially improved. With considerable publicity and declaration of a methodological revolution (Wagenmakers, 2015), psychological researchers discovered the need for preregistered well-powered confirmatory research that has long been the accepted practice for clinical trials.

However, psychological researchers are still struggling with methodological practices. Several key methodological issues have yet to be recognized and addressed. Psychological and parapsychological researchers can expect additional repetitions of this recent experience. When confronted with the choice between attributing experimental results to psi or to previously overlooked methodological deficiencies, psychological researchers will become advocates for additional methodological improvements.

An alternative strategy is to skip the inefficient cycles of implementing better methodological practices when skeptical psychologists become trapped by a choice between improved methodology and psi. Several methodological factors that psychologists and parapsychologists will eventually have to address can be easily foreseen and are discussed below. Methodological practices that address most of these factors have been implemented as standard procedure in regulated medical research. These practices are described.

Standards for Study Preregistration

Although the value of preregistration for confirmatory studies is now widely recognized, psychologists are still struggling with pitfalls in the registration process. Key practices for confirmatory

study registration include that: (a) the registry specifies requirements for the registration information, (b) all key methodological decisions that could affect the study outcome are registered, (c) each registration is independently reviewed for consistency and completeness, (d) the study registration is made irreversibly public before data collection begins, and (e) the registrations can be easily found and accessed when searching for studies on certain topics (Watt & Kennedy, 2015, also see the comments for that online article). The KPU Study Registry (2012) implements all of these practices, but, as yet, study registries for general use by psychologists do not incorporate all of these practices. Parapsychologists who use nonoptimal study registration can expect to have their findings eventually challenged when psychologists recognize the weaknesses of their registration practices.

Dropouts and Incomplete Data

Dropouts and incomplete data can introduce bias into experimental results but are often given little or no attention by researchers. Participants who are doing poorly in an experiment may tend not to complete the experiment or tend to make invalid responses. In general, dropouts and other types of incomplete data cannot be assumed to be independent of the experimental intervention and procedures—and therefore are confounding factors that can potentially cause bias. In addition, ad hoc retrospective decisions about handling incomplete data can create bias. The possibility that an ostensible finding is entirely or partially due to bias from incomplete data needs to be addressed.

In clinical trials the "intent-to-treat" (or "intention-to-treat") principle requires that every participant who was randomized is included in the primary analysis, whether or not the participant complied with the protocol (U.S. Food and Drug Administration, 1998; Gupta, 2011). Intent-to-treat is a fundamental principle guiding study conduct and analysis and is generally considered necessary to avoid bias. If the inclusion or exclusion of incomplete data alters whether the study outcome is significant, the incomplete data compromise the study and the results are unconvincing.

A secondary "per-protocol" analysis is often also done and includes only participants who substantially complied with the protocol. The per-protocol analysis may provide a better estimate of the effect size for participants who comply with instructions, but the analysis can also be severely biased if, for example, those with poor outcomes tend to drop out or tend to provide incomplete data.

Psychological and parapsychological researchers often discard incomplete data without fully considering the potential for bias. In effect, a per-protocol analysis is used as the primary analysis rather than an intent-to-treat analysis. This practice can promote bias in favor of large effects. Any experiment in which the participants receive real-time direct or indirect feedback that gives an indication of their performance is potentially subject to biased data-selection due to incomplete data. This is one example of the larger problem of confounding due to incomplete data.

A general principle is that the exclusion of any data for participants who began an experiment introduces potential for bias that must be carefully evaluated. This includes data removed as outliers. The handling of incomplete data frequently involves tradeoffs between potential biases that underestimate (intent-to-treat) or overestimate (per-protocol) the effect being investigated.

The intent-to-treat strategy is conservative and is generally optimal practice for providing evidence that an effect occurs. For nonmedical research, this principle might be called *intent-to-participate* and would be triggered when a participant has been randomly assigned to a group or has provided initial data for the effect being investigated. At a minimum, any excluded data should be analyzed and reported to evaluate possible bias. However, the key question is not whether the results for the excluded and included data are different, but whether the overall result is significant if the excluded data are included—which is the question answered by the intent-to-treat or intent-to-participate analysis. For missing data, sensitivity analysis exploring various unfavorable assumptions about the missing data should be conducted to evaluate possible bias. Useful guidance on these points can be found in U.S. Food and Drug Administration (1998). For confirmatory research the handling of incomplete data should be specified in the study preregistration and carefully explained in the study report.

Software Validation

Documented software validation is essential for confidence in research results. This principle is well established for regulated medical research (U.S. Food and Drug Administration, 2002, 2003) but has not yet become widely recognized for psychological and parapsychological research. Psychological researchers recognize the need to empirically evaluate the validity of questionnaires and other measurement instruments but have been slower to recognize the need to empirically evaluate the validity of other aspects of research methodology, such as software.

Software validation involves testing and documenting that the software reliably and accurately fulfills its purpose. Theoretical discussions of software validation can be highly technical and filled with jargon. However, a practical, common-sense approach is more realistic in practice.

Validation of the software used for data processing and analysis is straightforward. Another researcher develops and/or applies independent programs that duplicate the analyses.

Validation of the software used to conduct an automated experiment and to collect the raw data is more challenging. Verification that the software functions properly must be based on tests of the experimental procedure rather than on independent programming. Watt and Brady (2002) described a software oversight that produced artificial positive results in a psi experiment. Such oversights and errors can easily occur with automated research—but can be avoided with basic, established principles for software validation. In the absence of software validation, the possibility of artifacts due to programming errors and oversights is a legitimate concern.

End-user testing is the final and most important step for validating software. This step is performed by a user or tester who did not do programming for the study. End-user testing verifies that the software operates as intended for the specific environment that will be used in the experiment. This includes the specific version of the experimental software, the specific settings and options for the software, the specific computer operating system (including updates), the processor type and speed, storage devices, other devices attached to the computer, and other software that may be running in the background. Verification that the software functions properly in the experimental environment is particularly important if timing for the stimuli or responses has a significant role in the experiment.

Competent initial end-user testing often discovers problems that the software developer(s) did not anticipate. The key questions for validation of software for automated experiments are:

1. Does the software accurately and reliably present the stimuli and/or feedback for the experiment?
2. Does the software properly generate the random elements in the experiment?
3. Does the software accurately and reliably record the human inputs and the conditions generated by the software?
4. Does the software properly handle unexpected, inappropriate inputs?
5. Does the software have these properties for all computers that will be used in the experiment?

Software validation is based on documented empirical evidence that these questions are answered affirmatively, rather than relying on optimistic assumptions that the software and hardware operate ideally. A review of the programming source code by a knowledgeable person is usually valuable but does not replace the need for this empirical validation testing. The final version of the software that is used in an experiment, not just the programming language or system for developing the experimental software, needs to be validated.

End-user testing should detect intentional programming errors (fraud) as well as unintentional errors. The optimal practice is for the person who developed the software to not have access to the computers used for collecting data, and reciprocally, for the experimenters collecting data to not have access to the source code for the experimental software.

Methods and steps that may be useful for developing a plan for end-user testing are described in the

Appendix. These steps focus on testing the software used to conduct parapsychological and psychological experiments.

The development of scripts and datasets that generate known inputs and outputs is a common practice in validated software environments. The validation package is run when the software is initially installed on a computer and when changes are made, such as modification of the experimental software, changes or updates to the operating system, or changes to software that runs in the background. The validation package can also be run periodically to verify that the operation of the software has not been altered by factors unrecognized by the experimenters. These validation packages typically include automated comparison of the observed and expected output, and they provide a report of the validation results that is kept as part of the research records. Automated validation packages typically require significant effort to develop, but relatively little subsequent effort to apply.

Psychological researchers can be expected to recognize the need for software validation after some widely publicized cases of research retraction due to invalid software and/or after parapsychological findings force recognition of the possibility of unintentional or intentional programming errors.

Experimenter Fraud

If measures to prevent and to detect experimenter fraud are not implemented in research, fraud will often be easy and tempting with little possibility of detection. Typical research procedures in psychology and parapsychology do not include measures to prevent and to detect experimenter fraud. Stroebe, Postmes, and Spears (2012) noted that detected cases of fraud are likely the tip of an iceberg of mostly undetected fraud. In the absence of measures to prevent and to detect experimenter fraud, meaningful conclusions about the rate of occurrence of fraud are simply impossible (Kennedy, 2016c). The uncertainty about undetected fraud compromises confidence in research findings and is increasingly recognized as unacceptable. However, practical, effective measures to address experimenter fraud have not yet been widely recognized by psychological researchers.

Independent replication and peer review have not been effective at detecting and deterring experimenter fraud (Kennedy, 2014b, 2016c; Stroebe et al., 2012). The primary symptom of fraud is inconsistent results among experimenters, but such differences are virtually never attributed to fraud. Inconsistent experimental results in psychology and in parapsychology are typically attributed to differences in experimental procedures and subject populations. These alternative explanations could be true and prevent independent replication from being useful for detecting and deterring experimenter fraud. Differences among experimenters have been prominent throughout experimental parapsychological research and bring into focus the need to routinely use research methods that prevent fraud (Kennedy, 2014b, 2016c).

Experimenter fraud typically has been by an individual experimenter who had opportunity to manipulate or fabricate data with little chance of detection. I am not aware of any cases of fraud that involved collusion among experimenters in academic or nonprofit settings.

Experimental procedures that make undetected data changes or fabrications difficult for one experimenter were a methodological standard in my experience working in regulated medical research (Kennedy, 2014b; 2016c). This standard has also been recommended for parapsychology (Akers, 1984; Dalton et al.,1996; Kennedy, 2014b, 2016c; Rhine, 1974, 1975). The goal is to make undetected experimenter fraud difficult rather than easy and tempting. Procedures that involve duplicate records and experimenters checking each other are relatively easy to implement once these become standard practices for research.

One basic strategy is to send a duplicate copy of each component of the data to a secure location as early as possible in the data collection process. This should be done before an experimenter has unblinded information that could allow the experimenter to bias the results. E-mailing or uploading a copy to a distant person or website that serves as a secure data repository is good practice given modern technology. The optimal procedures will assure that a person is never alone with access to data that could allow biased manipulation without detection. For example, if an experimenter obtains the random number for the target for a trial from an online source, a second experimenter can observe the generation of the

target, the transmission of the target to the data repository, and the entry or recording of the target into the experimental database.

Automated experiments are not immune from experimenter fraud. A person with knowledge of programming for the software used to conduct the experiment could create a fraudulent version of the software that would be used covertly during the experiment. A version without the fraudulent programming would be used for control runs and for software validation. Or, the software could detect whether a run is an experimental or control run and have fraudulent bias apply only during experimental runs. Alternatively, the data file created by the experimental software could be modified by a separate program or by direct editing.

As noted above, the optimal practice is for the person who developed the software to not be involved with data collection and to not have access to the computer(s) used to conduct the experiments. The end user experimenters who collect the data should validate the software as described above to assure that the software operates as expected in the actual experimental environment. These experimenters should not have access to the source code for the software. If software is transferred from one experimenter group to another, the software should be validated on the computer(s) for the receiving experimenter group. The source code may also be transferred and reviewed, but here too, the experimenters managing the software and the experimenters collecting data should be different and have restricted access to the other functions. In addition, when possible, the automated software should send a copy of the data to a data repository before the output file could be modified. A duplicate archival copy of the software should also be kept at the repository and ideally a process implemented to verify that the software used for experiments was not modified from the original archival copy. Technical innovations for assuring the integrity of experimental software may be developed.

Making the raw data available to others for independent analyses is also a useful, but secondary, strategy for deterring and detecting fraud (Kennedy, 2014b, 2016c; Stroebe et al., 2012). Accusations of fraud based on post hoc analyses will too often be irresolvable given the intrinsic limitations of post hoc analyses. Stroebe et al. (2012) pointed out that such accusations will sometimes be incorrect due to the probabilistic nature of the analyses. In addition, data can be fabricated or altered in a way that does not leave convincing signs of fraud. Making the data available to others does not eliminate the need for procedures that prevent fraud.

Statistical Methods for Confirmatory Research

The criteria for determining whether a replication study is successful have received much attention recently as the need for preregistered, confirmatory research has come into focus (Lakens, 2016; Open Science Collaboration, 2015; Simonsohn, 2015a, 2016). Typically, one or two confirmatory studies have been done and must be compared with previous studies that had more questionable methodology.

In recent decades psychological researchers have excessively focused on *p* values without adequate consideration of effect size, statistical power, and the distinction between exploratory and confirmatory research. One of the most widely discussed reactions to this excess is Cumming's (2014) "new statistics," which focuses on estimating effect sizes and advocates abandoning hypothesis tests and associated "dichotomous thinking."

Good confirmatory research, like good science in general, is based on making and testing specific predictions. Exploratory research focuses on estimating effect sizes without specific predictions. Cumming's new statistics attempts to set exploratory research methods as the standard for scientific evidence and avoids the dichotomous thinking associated with testing predictions. That is a major retreat from the basic principles of good science. Researchers who apply Cumming's recommendations will tend to interpret experimental outcomes that can easily occur by chance as evidence for an effect and will explicitly avoid confronting the possibility that the effect being investigated may not be valid. The limitations of the new statistics are increasingly recognized (Kennedy, 2016b; Lakens, 2016; Morey, Rouder, Verhagen, & Wagenmakers, 2014; Savalei & Dunn, 2015).

Researchers conducting preregistered well-powered hypothesis tests evaluate specific predictions that provide the most convincing evidence that the researchers actually understand and control the effects being investigated. With the new statistics, psychologists have shifted from an excessive, misleading focus on *p* values to an excessive, misleading focus on effect sizes.

The basic challenge for evaluating confirmatory research is that *p* value, effect size, and statistical power must all be considered. No one parameter alone adequately conveys the full outcome. Although statistical power has been generally ignored by psychologists for decades, it is the heart of classical analysis and is a key component of the statistical validity of a hypothesis test (discussed below, also see Kennedy, 2016b). As Cohen (1990, p. 1310) commented, "failure to subject your research plans to power analysis is simply irrational." The statistical principles for regulated medical research clearly describe the importance of hypothesis tests, power, effect size estimates, and the distinction between confirmatory and exploratory research (U.S. Food and Drug Administration, 1998). These principles are useful guidance for anyone conducting experiments analyzed with statistics. Software such as the free program G*Power (Faul, Erdfelder, Lang, & Buchner, 2014) makes power analysis much easier than in the past.

One of the simplest and most informative presentations is the power curve or operating characteristics for a hypothesis test. This is a graph or table that shows the power of the test for different values of true effect size given the sample size in the study. This shows which effect sizes have low power as well as which have high power. For confirmatory research, the operating characteristics should be determined when the study is being planned.

As more experience is gained with power analysis, I expect that a statistical power of at least .90 (.95 when possible) will become the standard in situations when few confirmatory studies are expected, and in other cases when researchers want to use optimal methodology. A power of .80 may be appropriate when several or many confirmatory studies are expected and overall conclusions will be based on the results of multiple studies. The criterion for a successful confirmation will be a significant result in a study with a power of .80 or higher.

Properly designed confirmatory studies can provide evidence that the hypothesis of interest is false as well as true (Watt & Kennedy, 2015). If a study with a power of .90 or higher or multiple studies with a power of .80 fail to produce significant results, that outcome can be interpreted as evidence that the experimental hypothesis is false for the effect size used in the power analysis. The interpretation of nonsignificant results from studies with lower power or uncertain power is ambiguous because the results could be due to low power rather than to the experimental hypothesis being false.

A quantitative value for power is needed to provide evidence that an experimental hypothesis is false in the same way that a *p* value is needed to provide evidence that an experimental hypothesis is true. Without power analysis, a study cannot provide evidence that a hypothesis or prediction is false.

The effect size used in a power analysis when designing a study is a prediction about what will happen in the study. For a study with high power, a nonsignificant result provides evidence that the predicted effect size specified in the power analysis is false. A nonsignificant result does not provide direct evidence that the null hypothesis is true because a small, nonzero effect size could be true. The null hypothesis is basically irrelevant other than for developing a test for the effect predicted in the power analysis.

Researchers who abuse hypothesis tests focus on the null hypothesis and ignore statistical power. This prevents evidence that the experimental hypothesis is false and avoids the confirmatory question of whether the researchers can make reliable predictions about the effect being investigated. Most criticisms of hypothesis testing are actually criticisms of this abuse and are not applicable to proper applications of hypothesis tests with power analysis.

If a minimum effect size of interest cannot be reasonably specified for power analysis, the research is exploratory and the experimenters do not have the degree of understanding and control of the phenomenon that is needed for convincing evidence. Similarly, research is exploratory if the researchers ignore power analysis, as was common in the past in psychology and parapsychology. Exploratory research is typically the creative step that is the starting point for a line of research, whereas confirmatory research provides the convincing evidence that makes science valid and self-correcting.

If the power analysis and sample size for a confirmatory study are based on previous research, the uncertainty in the effect size estimates should be considered. As indicated by a confidence interval, the mean effect size from previous research can be assumed to have a 50% chance of overestimating the true effect size. The planned sample size should generally be based on an effect size on the low side of the confidence interval for the effects in the previous studies—such as the lower end of the 80% or 60% two-sided confidence interval (90% or 80% one-sided confidence interval). Allowance should also be made if the previous studies were exploratory and likely subject to bias.

Alternatively, the power analysis and sample size can be based on a theoretically meaningful effect size rather than on previous studies. For me an effect equivalent to a correlation of .14 or less (accounting for less than 2% of the variance) is basically meaningless in the social sciences except in the rare case that the effect has major health or economic implications. For many lines of research, an effect equivalent to a correlation of less than .2 (accounting for less than 4% of the variance) is inconsequential.

As discussed in Kennedy (2016a), studies with extremely high power (typically greater than .95) can give counterintuitive results if the *p* value for the outcome is significant but also is greater than one minus the power. This very rare situation is often pointed out by critics of classical hypothesis testing. The optimal practice in most cases with extremely high power is to set the alpha (significance) level for the study to be equal to one minus the power. Other options are discussed in Kennedy (2016a).

Multicenter studies can provide adequate power and optimal scientific evidence without placing a great burden on one research center. This is particularly important for parapsychological researchers who believe that prolonged testing by one experimenter group can cause the loss of the experimenter motivation and enthusiasm that is needed for successful psi results. Videos and other automated processes for interactions with participants can also alleviate concerns about declining experimenter motivation. The fact that such concerns raise serious questions about whether an ostensible psi effect is produced by the participants or by the experimenter is beyond the scope of the present paper.

Bayesian Analysis

Bayesian methods are becoming increasingly popular in psychology, but they are at a relatively early (honeymoon) stage of development. Bayesian methods currently have substantial uncertainties and potential pitfalls. These limitations will become widely recognized as experience is gained with these methods and as standards are developed (Kennedy, 2014a). One of the more prominent examples is that the common methods for Bayesian hypothesis tests have extreme sensitivity to the choice of a prior probability distribution (Kruschke, 2015, pp. 292–295, 346–348) and tend to be biased in favor of the null hypothesis, particularly for small effects and for diffuse prior probability distributions (Kennedy, 2015; Simonsohn, 2015b).

The development of operating characteristics for a statistical test as described above are needed for Bayesian hypothesis tests as well as for classical hypothesis tests (Kennedy, 2014a, 2015). The operating characteristics answer the fundamental question: If the true effect size is a certain value, what is the probability that the planned analysis will give the correct inference? These evaluations quantify the expected rates of inferential errors for a planned analysis and reveal unrecognized biases.

The evaluation of inferential errors establishes the statistical validity of a planned analysis. Statistical analysis is another component of research methodology that needs the validity of the planned methods to be evaluated and documented. Operating characteristics are expected when Bayesian hypothesis tests are used for regulated medical research (U.S. Food and Drug Administration, 2010).

Bayesian analysts who do not evaluate the expected rates of inferential errors for the predicted effect are continuing the statistical negligence that was common in the past among psychological researchers who applied classical statistics and similarly ignored power analysis. Psychological researchers have not yet widely recognized the need for these evaluations but are slowly moving in that direction. I expect that the debates about whether Bayesian or classical statistical methods are better will dissipate substantially when the statistical validity of the methods is quantitatively evaluated. In practice, rates of expected inferential errors are more important than philosophical ideas about the nature of probability.

Meta-Analysis

As discussed above, the appropriate standard for experimental evidence is that 80% or more of well-powered confirmatory studies produce significant results—which will usually make the occurrence of an effect obvious without a need for meta-analysis. Meta-analyses are most useful when adequately powered individual studies are not feasible or when deciding whether and how to pursue a line of research that has not yet obtained consistent confirmatory results.

A typical retrospective meta-analysis is similar to exploratory research because methodological decisions are made after the study outcomes are known. These decisions include which studies to include, what statistical methods to use, how to evaluate questionable research practices, and what moderating variables to investigate. These retrospective decisions provide opportunities for bias, and associated opportunities for challenging the results of the meta-analysis (Kennedy, 2013a).

A prospective meta-analysis is a form of confirmatory research because the methodological decisions are made before the included studies have been conducted (Watt & Kennedy, 2017). The statistical analysis plan for a prospective meta-analysis should be publicly preregistered before data collection begins for any of the included studies. Of course, the studies that may be included should also be individually preregistered. The registration for each study can be used to decide prospectively whether the study will be included in the subsequent meta-analysis. This process is described in the first such prospective meta-analysis initiated by Watt (2016; Watt & Kennedy, 2017). Prospective meta-analysis, with power analysis and adjustment for multiple analyses, will eventually replace retrospective meta-analysis as the preferred choice for controversial topics.

A prospective meta-analysis does not prevent or discourage research innovation or exploratory analyses during a meta-analysis. It simply prospectively specifies how a study will be handled for the preplanned analysis in a subsequent meta-analysis and clearly distinguishes between preplanned and post hoc analyses.

Statistical Dependence

Statistical dependence can be a problem in both parapsychological and psychological experiments. The basic issue is that human responses, both conscious responses and physiological measures, cannot be assumed to be independent, as is required for the dependent or outcome variable for standard statistical analyses (Kennedy, 2013b; 2014c). The effects of statistical dependence are difficult to predict, but a common effect is to make *p* values misleadingly small.

The traditional analysis for parapsychological experiments uses the random events as the outcome variable (Burdick & Kelly, 1977; Kennedy, 2014c). This strategy was selected specifically because the random events are independent (if properly generated) and thus avoid dependence problems. In general, an analysis that uses a *t* test or ANOVA with some type of human response as the outcome variable has potential for dependence problems. An analysis that uses a binomial or similar test with independent random events as the outcome variable is typically free of potential dependence problems.

The most severe dependence problems occur in studies with feedback to the participant on each trial, as occurs in presentiment studies that investigate physiological measures of precognitive anticipation of a random event (Kennedy, 2013b; 2014c). The participant's physiological responses may include reactions to the outcomes of previous trials and thus contain information about and be dependent upon the specific sequence of random targets or stimuli. This can introduce bias when the mean of the responses is found for each type of stimulus. Attempts to correct for dependence with statistical adjustments, or to do post hoc analyses to argue that certain types of dependence did not occur, are controversial (Kennedy, 2013b, 2014c). An alternative analysis that is consistent with the traditional strategy in parapsychology is to use the physiological measures to predict the random stimuli and use only data collected prior to feedback for a trial to make the prediction for the trial.

Bem's (2011) "feeling the future" experiments that had reaction time as the dependent variable are another case with potential dependence problems. However, these experiments are more complex than

the typical presentiment studies. Bem's studies involve random retroactive priming as well as random selection of stimuli, and also randomly intermixed forward (non-psi) priming trials. It is more difficult to imagine how bias from sequential dependencies could enter into the analyses of these data. However, it is also difficult to make a compelling argument that dependence problems cannot occur given the feedback on each trial.

Studies of remote influences on other persons (Schmidt, 2015) also have human responses as the outcome variable and thus have potential for dependence problems. However, these studies do not have feedback to the participant who makes the responses, which greatly reduces the potential for bias. These studies typically use the session as the unit of analysis rather than the epoch or trial, which further reduces the potential for dependence problems. Although I currently do not see potential dependence problems with these studies, the analyses are not as theoretically clean as analyses that use the random events as the outcome variable.

An alternative design and analysis for studies of remote influence has a session divided into sequential trials that each consist of two epochs. One of the epochs for each trial is randomly selected as the influence epoch and the other epoch is a control. The analysis is based on determining which of the two epochs for a trial has a response that is most consistent with the expected influence. That epoch is predicted to be the randomly selected influence epoch for the trial. A simple binomial analysis of the number of correctly predicted influence epochs would provide a theoretically clean analysis with no potential for dependence problems.

Experiments that have potential dependence problems place a cautious or skeptical scientist in the position of choosing between either psi or dependence problems as the most likely explanation for the results. Psi is unlikely to be the preferred choice. It is a safe prediction that skeptical psychologists will eventually interpret such psi studies as evidence for subtle dependence problems that need to be addressed. Parapsychological studies with the traditional analysis of random events as the outcome variable are much more difficult for scientists to dismiss.

Conclusions

The validity of all aspects of research methodology needs to be evaluated and documented for confirmatory research. This includes the validity of software and statistical methods as well as the validity of measurement instruments. Hypothesis tests that evaluate specific predictions about the experimental outcome provide the strongest evidence that researchers understand and control an effect. Quantitative evaluation of expected rates of inferential errors for the predicted effect sizes that the study is intended to detect is a fundamental component of statistical validity for both classical and Bayesian hypothesis tests. Procedures that prevent experimenter fraud are also an important component of valid research methodology, as is appropriate handling of incomplete data.

The methodological topics discussed here have not yet been widely recognized or addressed by psychological researchers. If parapsychological researchers adhere to the methodological practices of psychological researchers, they can expect that eventually skeptical psychologists will consider these methodological issues as more likely explanations for experimental results than psi. When confronted with this choice, the standards for research will be revised accordingly. In effect, this will repeat the cycle that has just occurred with the wide recognition that preregistered confirmatory research is needed. This inefficient process of methodological evolution can be easily avoided if psychological and parapsychological researchers proactively implement good research methodology.

References

Akers, C. (1984). Methodological criticisms of parapsychology. In S. Krippner & M. L. Carlson (Eds.), *Advances in parapsychological research 4* (pp. 112–164). Jefferson, NC: McFarland.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–425.

Burdick, D. S., & Kelly, E. F. (1977). Statistical methods in parapsychological research. In B. B. Wolman (Ed.), *Handbook of parapsychology* (pp. 81–130). New York, NY: Van Nostrand Reinhold.

Cardeña, E., Palmer, J. & Marcusson-Clavertz, D. (2015). *Parapsychology: A Handbook for the 21st Century,* Jefferson, NC: McFarland.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304–1312.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29. Retrieved from http://dx.doi.org/10.1177/0956797613504966

Dalton, K., Delanoy, D., Morris, R. L., Radin, D. I., Taylor, R., & Wiseman, R. (1996). Security measures in an automated ganzfeld system. *Journal of Parapsychology*, *60*, 129–147.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2014). G*Power. [Software]. Retrieved from http://www.gpower.hhu.de/

Gupta, S. K. (2011). Intention-to-treat concept: A review. *Perspectives in Clinical Research*, *2*, 109–112. Doi: https://dx.doi.org/10.4103/2229-3485.83221

Kennedy, J. E. (2013a). Can parapsychology move beyond the controversies of retrospective meta-analyses? *Journal of Parapsychology*, *77*, 21–35. Retrieved from http://jeksite.org/psi/jp13a.pdf

Kennedy, J. E.(2013b). Methodology for confirmatory experiments on physiological measures of precognitive anticipation. *Journal of Parapsychology*, *77*, 237–248. Retrieved from http://jeksite.org/psi/jp13b.pdf

Kennedy, J. E. (2014a). Bayesian and classical hypothesis testing: Practical differences for a controversial area of research. *Journal of Parapsychology*, *78*, 170–182. Retrieved from http://jeksite.org/psi/jp14.pdf

Kennedy, J. E. (2014b). Experimenter misconduct in parapsychology: Analysis manipulation and fraud. Retrieved from http://jeksite.org/psi/misconduct.pdf

Kennedy, J. E. (2014c) [Letter] *Journal of Parapsychology*, *78,* 273–274. Retrieved from http://jeksite.org/psi/jp14let.pdf

Kennedy, J. E. (2015). Beware of inferential errors and low power with Bayesian analyses: Power analysis is needed for confirmatory research. *Journal of Parapsychology*, *79,* 53–64. Retrieved from http://jeksite.org/psi/jp15.pdf

Kennedy, J. E. (2016a). Counterintuitive results for statistical tests with high power. Retrieved from http://www.jeksite.org/psi/high_power_p_value.pdf

Kennedy, J. E. (2016b). Critique of Cumming's "new statistics" for psychological research: A perspective from outside psychology. Retrieved from http://jeksite.org/psi/critique_new_stat.pdf

Kennedy, J. E. (2016c). Experimenter fraud: What are appropriate methodological standards? Manuscript submitted for publication. Retrieved from http://jeksite.org/psi/fraud_standards.pdf

KPU Study Registry. (2012). Retrieved from https://koestlerunit.wordpress.com/study-registry

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Burlington, MA: Academic Press.

Lakens, D. (2016). The statistical conclusions in Gilbert et al. (2016) are completely invalid. The 20% Statistician Blog. Retrieved from http://daniellakens.blogspot.co.uk/2016/03/the-statistical-conclusions-in-gilbert.html

Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E-J. (2014). Why hypothesis tests are essential for psychological science: A comment on Cumming (2014). *Psychological Science*, *25*, 1289–1290.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716 DOI: 10.1126/science.aac4716. Retrieved from http://datacolada.org/wp-content/uploads/2016/03/5341-Nosek-et-al-Science-2015-Estimating-the-reproducibility-of-psychological-science.pdf

Rhine, J. B. (1974). Security versus deception in parapsychology. *Journal of Parapsychology*, *38*, 99–121.

Rhine, J. B. (1975). Second report on a case of experimenter fraud. *Journal of Parapsychology*, *39*, 306–325.

Savalei, V., & Dunn, E. (2015). Is the call to abandon *p*-values the red herring of the replicability crisis? *Frontiers in Psychology*, *6*, 245. Retrieved from http://dx.doi.org/10.3389/fpsyg.2015.00245

Schmidt, S. (2015). Experimental research on distant intention phenomena. In E. Cardeña, J. Palmer, & D. Marcusson-Clavertz (Eds.), *Parapsychology: A Handbook for the 21st Century*. Jefferson, NC: McFarland.

Simonsohn, U. (2015a). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*, 559–569. Retrieved from http://datacolada.org/wp-content/uploads/2016/03/26-Pscyh-Science-Small-Telescopes-Evaluating-replication-results.pdf

Simonsohn, U. (2015b). The default Bayesian test is prejudiced against small effects. Data Colada Blog. Retrieved from http://datacolada.org/35

Simonsohn, U. (2016). Evaluating replications: 40% full ≠ 60% empty. Data Colada Blog. Retrieved from http://datacolada.org/47

Stroebe, W., Postmes, T., & Spears, R. (2012). Scientific misconduct and the myth of self-correction in science.

Perspectives on Psychological Science, *7*, 670–688. Retrieved from http://pps.sagepub.com/content/7/6/670.full.pdf+html

U.S. Food and Drug Administration, (1998). *Guidance for industry E9 statistical principles for clinical trials*. Retrieved from http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm073137.pdf

U.S. Food and Drug Administration, (2002). *General principles of software validation; Final guidance for industry and FDA staff*. Retrieved from http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm085371.pdf

U.S. Food and Drug Administration, (2003). *Guidance for industry Part 11, electronic records; electronic signatures — scope and application*. Retrieved from http://www.fda.gov/downloads/RegulatoryInformation/Guidances/ucm125125.pdf

U.S. Food and Drug Administration (2010). *Guidance on the use of Bayesian statistics in medical device clinical trials*. Retrieved from http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071121.pdf

Wagenmakers, E-J. (2015). A perfect storm: The record of a revolution. *The Inquisitive Mind*, Issue 25. Retrieved from http://www.in-mind.org/article/a-perfect-storm-the-record-of-a-revolution

Watt, C. (2016). A prospective meta-analysis of pre-registered ganzfeld ESP studies. Retrieved from http://www.koestler-parapsychology.psy.ed.ac.uk/Documents/KPU_Registry_1024.pdf

Watt, C., & Brady, C. (2002). Experimenter effects and the remote facilitation of attention focusing: Two studies and the discovery of an artifact. *Journal of Parapsychology*, *66*, 49–71.

Watt, C., & Kennedy, J. E. (2015). Lessons from the first two years of operating a study registry. *Frontiers in Psychology*, *7*, 173. Retrieved from http://dx.doi.org/10.3389/fpsyg.2015.00173

Watt, C., & Kennedy, J. E. (2017). Options for prospective meta-analysis and introduction of registration-based prospective meta-analysis. *Frontiers in Psychology, 7*, 2030. Retrieved from http://journal.frontiersin.org/article/10.3389/fpsyg.2016.02030/full

Broomfield, CO, USA
jek@jeksite.org

Appendix

Steps that may be useful in developing an end user validation plan for software used for collecting data are listed below. These steps are intended to give an idea of the possibilities and to stimulate thinking rather than to provide a fixed formula that is applicable in all situations.

Step 1. The validating experimenter or end user will initially familiarize himself or herself with the experimental software by acting as a participant for a few sessions.

Step 2. If the inputs to the program are from a keyboard or mouse, a keyboard and mouse recording program (such as JitBit Macro Recorder) will be set to capture keyboard and mouse events before the experimental software is run. (Some of these recording programs do not have the needed accuracy for the timing of inputs and playback—this should be tested.) If the experiment involves randomly selected displays, a video recording or video capture of the screen display (such as Replay Video Capture) will be used to capture the display. If timing is critical for the experiment, a video recording will often be preferable to software that captures the screen display (which utilizes significant computer resources and may interfere with the experimental program). If the software has settings that identify experimental runs versus control runs, the validation tests will be done with the settings for experimental runs.

A typical experimental session will be conducted by the validating experimenter. The data in the output file for the sessions will be verified by comparing it to the values captured by the keyboard, mouse, and display recordings. If a large amount of data is generated, it may not be feasible to check each data value. In these cases, it is important to verify the data for the first few trials and the last few trials in the experiment, and at the beginning and end of any runs or other subdivisions of the session. In addition, the data for randomly selected trials at other times can also be verified. If timing is important, video editing software can be used to examine individual frames and verify the timing. Also, the keyboard and mouse recording will include time between events, which may be useful information.

Step 3. If random processes are involved, as in a typical parapsychology experiment, the responses recorded for a typical session will be played back as input to the experimental software. This repetition of the session will be done at least 20 times to verify that the output appears to be random. The optimal practice when feasible would be to simulate the number of sessions in the planned experiment. If suspicious results occur, the sessions will be repeated until confidence is developed that the software either is appropriately random or is biased. The potential for interactions between the responses and random process can be evaluated by generating two sessions with extremes for the responses—such as one session with fast and/or biased responses and another session with slow and/or different biases. These sessions would be played back with the experimental software at least 20 times each (many more times if feasible) to verify random results. Unfortunately, if there is feedback for each trial and potential for biases from reactions to the feedback, the sessions testing for bias must be entered by a person.

Step 4. The validating experimenter will conduct a session that includes efforts to break the software. This includes a variety of unexpected responses, such as extremely rapid and premature responses, and slow and non-responses. Other tests include hitting extraneous keys, switching the Insert/Overwrite setting, switching the Caps Lock setting, holding down a key to make the input repeat, and exploring various other unexpected inputs. Any unexpected behavior of the software will be noted and the data output will be compared with the recordings of the keyboard and mouse events, and with the video of the display if relevant.

Step 5. In addition to these initial validation steps, testing can be done based on the results obtained during the experiment. For the sessions with the most extreme effects in the actual experiment, the responses and timing for the responses can be played back with the experimental software at least 20 times (more if feasible) to verify that the results are generally random. The optimal strategy would be to use an independent program to capture and play back the keystrokes and mouse events for all experimental sessions. However, if necessary the scripts for the playback could be developed from the data collected for the experiment.

Abstracts in Other Languages

German

IST DIE METHODOLOGISCHE REVOLUTION IN DER PSYCHOLOGIE VORUEBER ODER BEGINNT SIE GERADE ERST?

ZUSAMMENFASSUNG: Signifikante Ergebnisse parapsychologischer Experimente, die unter Verwendung üblicher psychologischer Forschungsmethoden erzielt wurden, haben Psychologen dazu veranlasst, einige weitverbreitete methodologische Mängel und die Notwendigkeit vorher registrierter konfirmatorischer Forschung von zufriedenstellender Teststärke anzuerkennen. Psychologische Forscher sehen jedoch noch nicht weitere allgemein verbreitete methodologische Mängel ein, von denen man erwarten kann, dass sie zum Anlass der Wiederholung dieses Zyklus‘ werden. Vor die Wahl gestellt, sich zwischen Psi und unentdeckt gebliebenen methodologischen Mängeln entscheiden zu müssen, werden Psychologen die Notwendigkeit methodologischer Verbesserungen vorziehen. Dazu zählen folgende übersehenen methodologischen Faktoren: (a) fehlerhaftes Vorgehen bei der Studienregistrierung, (b) Verzerrungen aufgrund von Abbrechern und unvollständigen Daten, (c) die Notwendigkeit einer Software-Überprüfung, (d) Maßnahmen zur Verhinderung von Experimentatorbetrug, (e) angemessene statistische Methoden zur konfirmatorischen Forschung, (f) das Versäumnis, mittels Bayesscher Analysen auf fehlerhafte Schlußfolgerungen zu achten, (g) die Mängel einer retrospektiven Metaanalyse und die Vorteile einer prospektiven Metanalyse und (h) Probleme aufgrund der statistischen Abhängigkeit der Ergebnisvariablen bei statistischen Analysen. Psychologische und parapsychologische Forscher können diesen ineffizienten Prozess einer methodologischen Evolution unter Berücksichtigung von Kontroversen über parapsychologische Ergebnisse leicht vermeiden. Forschungspraktiken, die diese methodologischen Mängel berücksichtigen, stehen zur Verfügung, und ihre Nützlichkeit für psychologische und parapsychologische Forschung wird sich am Ende erweisen. Die empfohlenen Praktiken zur Vermeidung dieser methodologischen Mängel werden beschrieben.

Spanish

¿HA TERMINADO LA REVOLUCIÓN METODOLÓGICA EN LA PSICOLOGÍA O APENAS PRINCIPIA?

RESUMEN: Los resultados significativos en experimentos parapsicológicos utilizando métodos de investigación psicológica estándar motivaron a los psicólogos a reconocer deficiencias metodológicas generalizadas y la necesidad de pre-registrar investigación confirmativa con suficiente poder. Los investigadores psicológicos aún no han reconocido varias otras debilidades metodológicas comunes que pueden hacer que este ciclo se repita. Cuando se enfrenten a la elección entre aceptar fenåomenos psi o corregir deficiencias metodológicas pasadas por alto, los psicólogos reconocerán la necesidad de mejoras metodológicas. Estos factores metodológicos no contemplados incluyen: (a) prácticas deficientes de registro de estudios, (b) sesgo por abandonos y datos incompletos, (c) la necesidad de validación de los programas de computación, (d) medidas para prevenir fraudes por los experimentadores, (e) métodos estadísticos apropiados para confirmación, (f) no considerar errores inferenciales con análisis bayesianos, (g) las debilidades del metanálisis retrospectivo y las ventajas del metaanálisis prospectivo, y (h) los problemas de dependencia estadística en las variables de resultado en los análisis estadísticos. Los investigadores psicológicos y parapsicológicos pueden fácilmente evitar este ineficiente proceso de evolución metodológica causados por las controversias sobre los hallazgos parapsicológicos. Las prácticas de investigación diseñadas para resolver estas deficiencias metodológicas están disponibles y serán reconocidas como necesarias para la investigación psicológica y parapsicológica. Se describen prácticas recomendadas para abordar estas debilidades metodológicas.

French

EST-CE LA FIN OU LE COMMENCEMENT DE LA RÉVOLUTION MÉTHODOLOGIQUE EN PSYCHOLOGIE ?

RESUME : Des résultats significatifs obtenus lors d'expérimentations parapsychologiques utilisant des méthodes de recherche standards en psychologie ont poussé les psychologues à reconnaître des déficiences méthodologiques largement diffusées, et le besoin pour une recherche confirmatoire pré-enregistrée et dotée d'une puissance statistique suffisante. Les chercheurs en psychologie n'ont pas encore reconnu d'autres faiblesses méthodologiques communes qui pourront, selon toute attente, entraîner une répétition de ce cycle critique. Lorsqu'ils sont confrontés au choix entre "psi" et "déficiences méthodologiques sous-estimées", les psychologues vont reconnaître le besoin d'améliorations méthodologiques. Ces facteurs méthodologiques sous-estimés incluent : (a) pratiques déficientes en enregistrement d'étude, (b) biais liés à des données exclues ou incomplètes, (c) le besoin d'une validation de software, (d) des mesures pour prévenir la fraude expérimentale, (e) des méthodes statistiques appropriées pour la recherche confirmatoire, (f) l'incapacité à considérer les erreurs inférentielles avec les analyses bayésiennes, (g) les faiblesses des méta-analyses rétrospectives et les forces des méta-analyses prospectives, et (h) les problèmes de dépendance statistique pour les variables indiquant les résultats lors d'analyses statistiques. Les chercheurs en psychologie et parapsychologie peuvent facilement éviter ce processus inefficace d'évolution méthodologique contrainte par des controverses autour des résultats de la parapsychologie. Des pratiques de recherche qui font face à ces déficiences méthodologiques sont déjà disponibles et seront probablement reconnues comme nécessaires pour la recherche en psychologie et en parapsychologie. Nous décrivons ces pratiques recommandées pour contrer ces faiblesses méthodologiques.