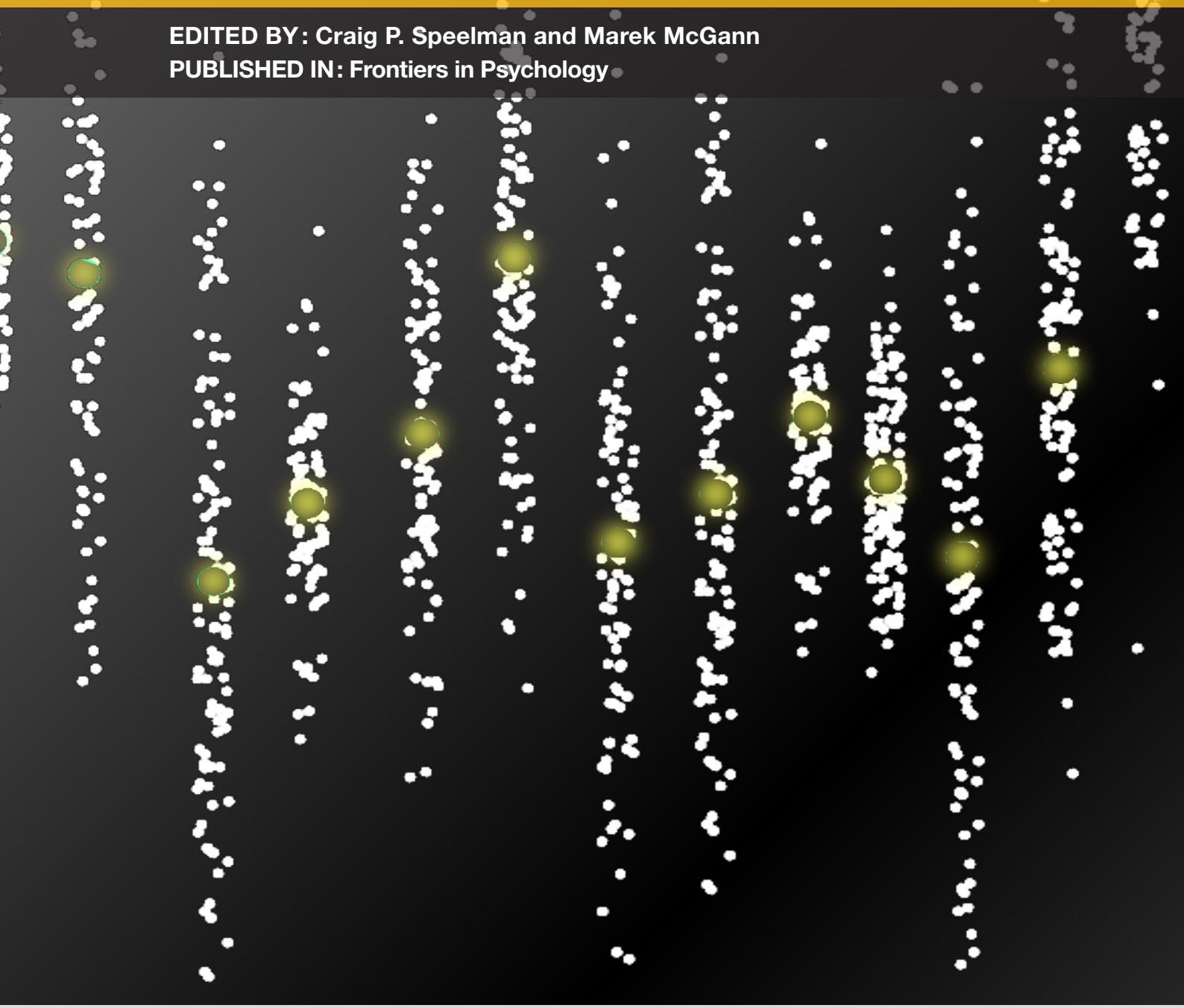
CHALLENGES TO MEAN-BASED ANALYSIS IN PSYCHOLOGY: THE CONTRAST BETWEEN INDIVIDUAL PEOPLE AND GENERAL SCIENCE

EDITED BY: Craig P. Speelman and Marek McGann
PUBLISHED IN: Frontiers in Psychology



frontiers Research Topics

frontiers

Frontiers Copyright Statement

© Copyright 2007-2016 Frontiers Media SA. All rights reserved.

All content included on this site, such as text, graphics, logos, button icons, images, video/audio clips, downloads, data compilations and software, is the property of or is licensed to Frontiers Media SA ("Frontiers") or its licensees and/or subcontractors. The copyright in the text of individual articles is the property of their respective authors, subject to a license granted to Frontiers.

The compilation of articles constituting this e-book, wherever published, as well as the compilation of all other content on this site, is the exclusive property of Frontiers. For the conditions for downloading and copying of e-books from Frontiers' website, please see the Terms for Website Use. If purchasing Frontiers e-books from other websites or sources, the conditions of the website concerned apply.

Images and graphics not forming part of user-contributed materials may not be downloaded or copied without permission.

Individual articles may be downloaded and reproduced in accordance with the principles of the CC-BY licence subject to any copyright or other notices. They may not be re-sold as an e-book.

As author or other contributor you grant a CC-BY licence to others to reproduce your articles, including any graphics and third-party materials supplied by you, in accordance with the Conditions for Website Use and subject to any copyright notices which you include in connection with your articles and materials.

All copyright, and all rights therein, are protected by national and international copyright laws.

The above represents a summary only. For the full conditions see the Conditions for Authors and the Conditions for Website Use.

ISSN 1664-8714
ISBN 978-2-88945-043-5
DOI 10.3389/978-2-88945-043-5

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

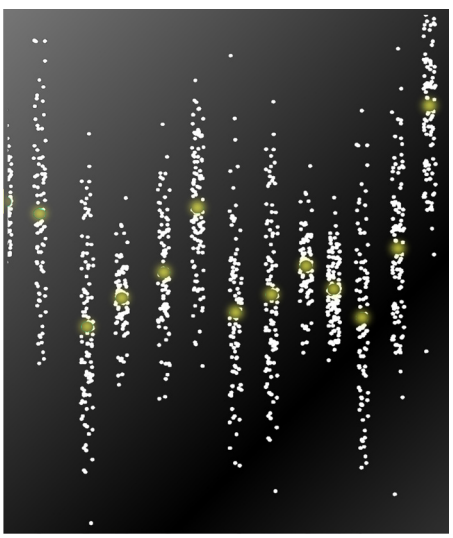
Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: **researchtopics@frontiersin.org**

CHALLENGES TO MEAN-BASED ANALYSIS IN PSYCHOLOGY: THE CONTRAST BETWEEN INDIVIDUAL PEOPLE AND GENERAL SCIENCE

Topic Editors:
Craig P. Speelman, Edith Cowan University, Australia
Marek McGann, Mary Immaculate College, University of Limerick, Ireland



Cover image by Marek McGann

In a recent paper we (Speelman & McGann, 2013) argued that psychology's reliance on data analysis methods that are based on group averages has resulted in a science of group phenomena that may be misleading about the nature of and reasons for individual behaviour. The paper highlighted a tension between a science in search of general laws on the one hand, and the individual, variable, and diverse nature of human behaviour on the other. This Research Topic explored this concern about the pitfalls of using the mean for the basis of psychological science. The problem is universal in its applicability to psychology, and opinion papers, reviews, and original empirical research from all areas of the discipline were invited. A total of 16 authors contributed 9 articles to the Topic. The range of issues that the authors viewed through the lens provided is impressive. The papers in this collection include a range of perspectives that provide concrete examples of how to approach research design, data collection, and analysis differently. No one contribution will provide a solution to our multifarious challenges, but nor should it. Our subject matter is complex and subtle, our investigations and methodological techniques will need to be equally so.

Citation: Speelman, C. P., McGann, M., eds. (2016). Challenges to Mean-Based Analysis in Psychology: The Contrast Between Individual People and General Science. Lausanne: Frontiers Media. doi: 10.3389/978-2-88945-043-5

Table of Contents

frontiers in
PSYCHOLOGY

REVIEW ARTICLE
published: 23 July 2013
doi: 10.3389/fpsyg.2013.00451

How mean is the mean?

Craig P. Speelman [1]* ***and Marek McGann*** [2]

[1] School of Psychology and Social Science, Edith Cowan University, Joondalup, WA, Australia
[2] Department of Psychology, Mary Immaculate College, University of Limerick, Limerick, Republic of Ireland

Edited by:
Lisa Lix, University of Saskatachewan, Canada

Reviewed by:
Yanyan Sheng, Southern Illinois University, USA
Fernando Marmolejo-Ramos, Univerity of Adelaide, Australia

***Correspondence:**
Craig P. Speelman, School of Psychology and Social Science, Edith Cowan University, 270 Joondalup Drive, Joondalup, WA 6027, Australia
e-mail: c.speelman@ecu.edu.au

In this paper we voice concerns about the uncritical manner in which the mean is often used as a summary statistic in psychological research. We identify a number of implicit assumptions underlying the use of the mean and argue that the fragility of these assumptions should be more carefully considered. We examine some of the ways in which the potential violation of these assumptions can lead us into significant theoretical and methodological error. Illustrations of alternative models of research already extant within Psychology are used to explore methods of research less mean-dependent and suggest that a critical assessment of the assumptions underlying its use in research play a more explicit role in the process of study design and review.

Keywords: mean, average, variability, noise, distributional analyses, cognition

INTRODUCTION

Psychology prides itself on its empirical basis. All undergraduate psychology courses focus on training students in methods for collecting and analyzing data about human behavior. To this end, a common introductory lesson in psychology involves measuring a group of humans on some variable and calculating the mean of the values. This mean value is then discussed as representing the average performance of the group, as if this value provides a representative substitute for the group's data. It is our view that rationalizing a set of data into one value is a theoretically loaded practice that can be misleading and possibly erroneous. While the mathematical tool itself is theory neutral, its use within the community of scientific practice within psychology is not. In this paper we argue that the mean, used without care, can cause illusions of stability and reliability in behavioral data, which in turn leads to inappropriate conclusions regarding the underlying nature of the psychological system. Our principal intent in what follows is to make the assumptions inherent in typical scientific practices more explicit, to expose them for critique. Our examination of the mean is therefore less a statistical one than a theoretical one. Having articulated these concerns we explore a number of related alternative theoretical perspectives that are less reliant on those explicated assumptions. We argue that, at the very least, we should take more care in our use of the mean in analyzing data. Better yet would be the adoption of methods and theoretical frameworks that cope better with the complexity and variability of behavior and cognition.

THE MEAN AND ITS USAGE

The most common form of the mean that is used in psychology is the arithmetic mean. This represents the sum of all of the values in a set divided by the number of values in the set. Although other forms of the mean are used in psychology (e.g., geometric, harmonic), our arguments are mainly confined here to the arithmetic mean. Our concerns, though, are mainly with the use of measures of central tendency, so our arguments apply in the general sense to all forms of the mean.

Textbooks used in introductory psychology courses on statistics and research methods typically refer to the mean as a measure of central tendency and often contrast it to other measures of central tendency such as the mode and the median. All such measures are used in situations where data sets contain some variation (i.e., not every score has the same value). The aim in calculating one of these measures, then, is to generate a value that sits somewhere in the middle of the distribution of scores. On the relatively rare occasions that the particular importance of measures of central tendency is mentioned it is as an indicator of the typical or most likely scores within the distribution (e.g., Haslam and McGarty, 2003, p. 135; Dancey and Reidy, 2008, p. 44; Howitt and Cramer, 2011, p. 25). These "typical" scores are "summary" or "descriptive" statistics, providing at least some insight into the basic characteristics of the distribution in question.

However, while all of the data in a distribution are involved in calculation of the arithmetic mean, it remains a matter of judicious use as to how well the mean *represents* those data. Summaries are vital to good communication but used too frequently and uncritically they provide an impression of reliability or consistency that distorts the normal state of affairs. It is this sometimes careless overuse of the mean and its too frequent use without other statistics as a summary of distributions with which we are concerned here. In particular, the over-reliance on the mean (despite the fact that we all, of course, know better) expresses a way of thinking about distributions and variability that we believe poses potentially grave problems for our science. Introductory textbooks typically indicate that most measures of anything related to humans, and indeed any biological system, produce a distribution of scores. These distributions are commonly normal in shape (although see Micceri, 1989). That is, they have a shape that corresponds to the Gaussian distribution, which has a symmetrical shape, with most scores clustered around the middle of the distribution, fewer scores at the tails, and a smooth transition from the middle to the tails (see **Figure 1**). The mean, therefore, sits perfectly in the middle of the normal distribution (as does the median and the mode).

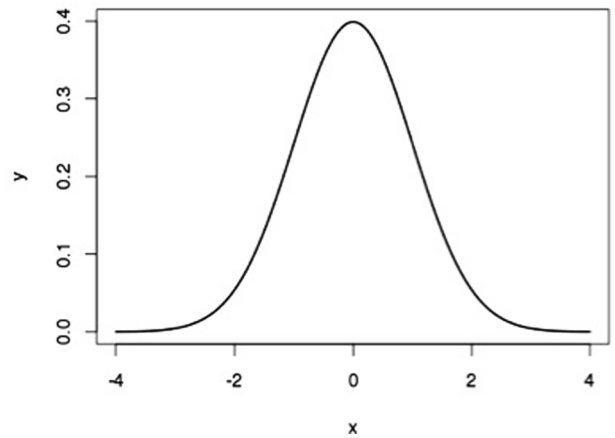


FIGURE 1 | The Gaussian, or "normal," distribution. *X* represents the measured variable, and Y represents the probability of occurrence of particular values of *X*.

The Gaussian distribution is also referred to as the "normal law of error" (Boring, 1920), suggesting that scores on either side of the mean represent some error of measurement. Adolph Quetelet, the first to apply this law to social and biological data, suggested that the mean of a distribution of human measurements, such as of a set of heights, represented nature's ideal value and that values on either side of the mean were deviations from nature's ideal (Howell, 2002)[1].

Relying on the mean of a set of scores to represent the set appears to carry with it the assumption that the variation in values observed around the mean is somehow erroneous. Whether it be that our methods of measurement are faulty, or that individual humans represent "deviations around nature's ideal," or both, values around a mean are often considered noise, and the only way to eliminate this noise is to average it away. What we are left with, then, is an approximation to the "true" value for that human dimension. It is this assumption regarding the interpretation of noise, and the truth value of the mean, that we think requires questioning. This assumption has several subsidiary assumptions (see **Table 1**) that we tackle below. In the sections that follow we review each of these assumptions and conclude that they are difficult to justify. So too are some of the implications of using the mean to infer features of the human cognitive system.

The uncritical or unreflective use of the mean in much psychological research makes us problematically blind to variation and distribution amongst the data we collect. In focusing narrowly on the mean we make ourselves blind to potential variation and complexity in our data and in the cognition and behavior those data represent. The assumptions we identify as underlying the uncritical use of the mean are, effectively, assumptions about certain characteristics of the psychological system, characteristics that tell us more about the theoretical goggles we are wearing than about the behavior we are observing. In later sections of the paper we consider alternative approaches that help us shake these assumptions, and suggest how these may provide fruitful means of conducting research in psychology.

Table 1 | Assumptions underlying the use of the mean in psychology research.

1. There is a true value that we are trying to approximate when we measure humans on some dimension.
2. Averaging helps us to eliminate the noise in our measures to see the true value.
3. Any inability to use the mean as a reliable measure of a stable characteristic is a product of weaknesses in methodology or calculation (i.e., it does not represent a failure in the initial assumption that a true value exists).
4. The noise in our measurements represents the effects of variables unrelated to the one being measured.

ASSUMPTION 1: THERE IS A TRUE VALUE THAT WE ARE TRYING TO APPROXIMATE WHEN WE MEASURE HUMANS ON SOME DIMENSION

What is it about human behavior we are trying to eliminate by averaging? In our own field of cognitive psychology it would seem we assume that in each head there is a mechanism that is common to all/most people, but which is obscured by our noisy measures and/or our noisy heads. That is, in our experiments, we expose a group of people to the same conditions. Everyone is assumed to respond similarly to these conditions because their cognitive mechanisms are similar. Unfortunately the data we collect from these people are not identical, and we assume this is because our measurements are not perfect and that there are a myriad of tiny and random effects that conspire to create noise in the data. Still, if we test a sufficiently large sample size, averaging should enable us to observe the characteristics of each cognitive mechanism unobscured by the noise.

The main question that occurs to us when we consider this scenario is why do we assume that everyone has the same cognitive mechanism? Just as we would not readily accept that each person's height is some deviation from an ideal height, it is odd that we would accept that each person's brain works in exactly the same manner. Certainly this is the assumption that our research methods in cognitive psychology rest upon, and yet there does not appear to be any attempt to justify it[2].

One means of justifying this assumption could be to point to other systems in the human body and note that they all tend to

[1]Other distributions are also observed with respect to measurements of human characteristics, but these often reflect something peculiar to the characteristic under scrutiny. For example, log-normal distributions are predominant in situations involving some degree of competition or interaction between the elements being measured (e.g., Halloy, 1998). Reaction time distributions (seen as ex-Gaussian distributions) are invariably positively skewed, and this has inspired much model development and testing (e.g., Ratcliff and McKoon, 2008; Holden et al., 2009; Heathcote and Love, 2012).

[2]This is the strong version of the assumption. The weak version is that we assume that people have *pretty much the same* cognitive mechanism. This, however, is not a version that helps. If we accept that people have slightly varying cognitive mechanisms, using the mean to draw inferences about these mechanisms – that is, to get a picture of some average mechanism – we end up with a mechanism that may not exist in anyone's head. More on this later.

work in similar ways in each individual. For example, the heart operates in the same manner in each person and although some viable deviations from the standard exist (e.g., atrial septal defect, dextrocardia), the vast majority of people have similar cardio-vascular systems. Most of the other major systems in the body also have the same uniformity across the human species, from the cellular to musculo-skeletal levels. In response to this justification, however, we would point to the fact that the brain has one major difference to the other systems in the body – it changes its mode of operation as a function of experience[3]. We consider this response further below, but for now we suggest that the assumption of common cognitive mechanisms is one that can be challenged, and probably should not be the starting point in explaining human behavior.

ASSUMPTION 2: AVERAGING HELPS US TO ELIMINATE THE NOISE IN OUR MEASURES TO SEE THE TRUE VALUE

We have no issue with the common statistical notion of sampling error. This is the notion that, when sampling from a population of scores, each sample will have a mean that is likely to vary from the mean of the population with a fairly predictable probability. That is, there are likely to be many samples with means that fall fairly close to the population mean, and a much smaller number that have means further away from the population mean. The chances of obtaining a sample mean close to the population mean are increased by taking a larger sample.

The problem we have with sampling is more in the interpretation of sampling error. Just as we have a difficulty with the concept of the mean reflecting a true value on some variable, we also find it challenging to accept that sample values on either side of the mean reflect noise in the data. This interpretation suggests that these values are not psychologically meaningful. Instead they are a nuisance factor that requires elimination. Indeed, if this noise did not exist, if we could measure "true" values directly, we would have no need for inferential statistics such as the analysis of variance (ANOVA).

Despite there being something of a tradition within psychology pointing out the difficulty in this assumption (see particularly recent consideration by Doherty et al., 2013, and more classically, Meehl, 1978), standard practice would appear to hold tight to this assumption. Just as there are no obvious justifications for the argument that the mean reflects some true value, we have discovered no explicit attempts to justify the elimination of variance as a "cleaning-up" activity. It just appears to be the done thing.

A crucial mathematical (as opposed to psychological/theoretical) assumption regarding the use of the average to eliminate noise is the shape of the distribution in question. For the assumed Gaussian curve averaging provides us with a clear representation of the center of the distribution, the "noise" to either side being averaged away. In a sobering and landmark paper, Micceri (1989) noted that normal curves are very rare in real psychological data. To calculate the mean in the hope of eliminating noise or getting some glimpse of a "true" or even a typical value hidden in the variation is simply to overlook reality in favor of a comfortingly elegant mathematical ideal.

ASSUMPTION 3: ANY INABILITY TO USE THE MEAN AS A RELIABLE MEASURE OF A STABLE CHARACTERISTIC IS A PRODUCT OF WEAKNESSES IN METHODOLOGY OR CALCULATION (i.e., IT DOES NOT REPRESENT A FAILURE IN THE INITIAL ASSUMPTION THAT A TRUE VALUE EXISTS)

One of the assumptions underlying averaging is that our methodologies are inherently faulty in that they cannot be expected to provide perfect measures of the variables of interest. To some extent, this assumption is indisputable, considering that even measurements of physical properties (e.g., length, temperature) carry with them conventional measurement error values. In psychology, however, we take this assumption further than in the physical sciences. Although we accept that there are features of the physical environment that will affect the accuracy of any measurements we take, we are also concerned with the validity and the reliability of the measures. Validity reflects whether we are measuring what we think we are measuring. Most often psychological variables are not directly observable so we need to construct measures that are directly observable and argue that these reflect the operation of the unobservable mechanisms we are interested in. Even if we assume that our measures are valid in this sense, the reliability of these measures concerns psychology greatly. Indeed, Psychological Test Theory makes explicit this notion by indicating that each score on a particular test reflects the true value for that person on the test, plus error (Novick, 1966). In cognitive psychology, we do not seem to believe that our measures are capable of producing an accurate reflection of the state of a person's cognitive system at some point in time. Indeed, if we exposed a person to the same stimuli, under the same conditions, on several occasions, and recorded their reaction times in responding to those stimuli, we would likely average the individual reaction times, on the assumption that each RT could not reflect the "true" RT for that person in that condition.

There is little doubt that there would be variance amongst the RT values recorded in this situation but what is the justification for assuming that each RT is a deviant of the true value? Instead of assuming that each RT is the true value plus some error created by seemingly random processes, could it not be possible that each RT reflects the state of the cognitive system *as it is at that point in time*? By this we mean, behavior in response to the experimental conditions reflects not only the external conditions, but also the state of the cognitive system as the behavior is occurring. The system will be in a different state to the one it was in on the previous occasion when an RT was recorded, and to the state when the next RT is recorded, if for no other reason than the fact that the system has experienced a repetition of the experimental conditions and made the same responses. The assumption then that taking the average of measures from repeated trials will provide a reflection of some stable element of the cognitive system seems fanciful given that the system could not be stable if we keep giving it experiences. This is a psychological reflection of the Heisenberg Uncertainty Principle – by measuring a system, we are influencing the system and hence affecting the very thing we are trying to measure. Unfortunately,

[3]Other systems in the body do this also (e.g., the cardio-vascular system responds to regular exercise by becoming more efficient), but this is usually in a quantitative manner, whereas the brain alters its functioning in not only a quantitative, but also a qualitative manner (Patel et al., 2013).

averaging several scores will not result in a value that reflects some stable feature of the cognitive system. Instead, this stable feature may not even exist.

ASSUMPTION 4: THE NOISE IN OUR MEASUREMENTS REPRESENTS THE EFFECTS OF VARIABLES UNRELATED TO THE ONE BEING MEASURED

Part of the justification for averaging scores appears to be the assumption that scores on either side of the mean reflect error. This error can be error of measurement, as considered above, but it can also reflect the operation of many factors that influence behavior. There is recognition in psychology that humans are sensitive to a vast range of variables, and any measurement of one variable is going to show the effects of many of these other variables too (hence the widespread usage in psychology of statistical methods such as Factor Analysis and Structural Equation Modeling). However, we assume that these effects have several characteristics. One is that they are random, and the other is that they operate independently of the variable we are interested in (i.e., the one we are currently measuring). Essentially, then, this error is assumed to be analogous to white noise in a radio signal. As such, calculating the mean is assumed to be just like fine tuning a radio signal – in both situations, noise is eliminated to enable a clearer perception of the signal. Again, though, we wonder what justification there is for such an assumption. Below we consider whether it is appropriate to assume that the variance in a set of scores is a reflection of variables that are random and independent of the variable we have measured.

INTERPRETIVE CONSEQUENCES OF USING THE MEAN

The mean is often used as if it is a good representation of a group of scores. Clearly it is only used when there is variation amongst scores – if there is no variation, then the scores can be characterized easily as so many scores of the same value. When there is variation amongst the scores, some measure that reflects the middle of the distribution of scores is considered to be a good reflection of the type of score that is observed in that set. As this variation in the set increases, however, the confidence one has in the mean being a good reflection of the group decreases.

This raises an issue with respect to the testing of differences between groups in an experimental design. Standard inferential statistics compare the variation between groups with the variation within groups in order to determine whether the scores in one group are significantly different to those in the other group. In spite of its name, then, ANOVA is rarely used to determine whether differences exist between the variances of two or more samples. Although there is explicit recognition that it is variance we are considering, ultimately the conclusions that are drawn in such situations concern whether the mean of one group is different to the mean of the other group. So, even though the statistical test explicitly considers the degree of overlap between the distributions of scores in the groups, the final conclusion is phrased in terms of whether one mean is significantly larger than the other.

The extent to which such a conclusion is a fair reflection of the state of the distributions is of course affected by the size of the difference between the means but it is also influenced by the amount of variation in the distributions, and the number of scores in each distribution. Thus it is not uncommon to see significant differences between means reported where the differences are very small. This will happen when the variation within groups of scores is small, and/or the number of scores is large (Cumming, 2012, chapter 12).

In the end, though, a conclusion that the mean of one group is different to the mean of another group can end up being translated as one condition improved performance more than another condition, or similar. But is this an accurate summation of the outcome of the statistical test, and for what purposes are such conclusions used? When one concludes that one condition led to better performance than another condition, at best the implicit conclusion is that, on average, or generally, this condition improves performance. But, as is clear from the above characterization of inferential statistics, there may well be significant overlap in the scores between the two conditions. For instance, if Condition A led to better overall performance than Condition B, there could well be scores in Condition B that were better than scores in Condition A. In other words, the final conclusion may represent an accurate description of the state of affairs for a subset of scores, but not necessarily for the whole set. Many undergraduate statistical courses and textbooks include such caveats in the discussion of the outcome of statistical tests, but the practices of scientists and standards of review and publication in journals involve few checks or balances against this kind of concern (see Marmolejo-Ramos and Matsunaga, 2009 for examples and explorations of good practice in this regard).

Inferential statistical tests generally do not provide information regarding the number of cases that do or do not match the pattern of results represented by the difference in means[4]. It is, of course, a trivial matter to generate such information. Doing so can provide illuminating results.

For example, one of us teaches a unit in Cognition in which one laboratory exercise involves replicating the Word Superiority Effect. This is where detection of a letter is found to be more accurate when the letter is presented in the context of a word than when presented in isolation (Reicher, 1969; Wheeler, 1970). Data has been collected in this laboratory exercise for over 5 years. Although the standard word superiority result is found with these data and supported by a statistically significant superiority in the word condition, when individual scores are considered, almost half of the over 500 people in the experiment provided results that either showed no difference between the conditions (i.e., identical accuracy scores in each condition), or showed results that were the opposite of the effect. Although there may well be methodological differences between our experiment and the classic versions published by Reicher and Wheeler, this observation does raise a serious question over the validity of using inferential statistics to assess differences between means. If we just examine the differences between means and focus only on whether or not this difference is statistically significant, we can end up with a conclusion that describes the effect of the manipulation as if it has had the effect on all or most of the individual scores in the data set. In other words, the mean difference can ultimately represent all

[4]See Cohen (1977) for a measure (U3) that estimates such group differences *post hoc*.

Table 2 | Number (%) of empirical articles in *Memory & Cognition* (2012) and *Journal of Experimental Psychology: Learning, Memory, and Cognition* (2012) classified according to main analysis type.

Journal	$\bar{X}$/NHST	Ind. Diffs	Other	Total
M&C	88 (82.2%)	17 (15.9%)	2 (1.9%)	107
JEP: LMC	79 (81.44%)	6 (6.19%)	12 (12.37%)	97

$\bar{X}$/NHST, summary statistics (e.g., mean, frequencies) and null hypothesis significance testing; Ind. Diffs, analysis explored individual differences in responses; Other, analysis could not be classified into one of the other two categories (e.g., structural equation modeling, chi-squared, Bayesian analyses of various kinds, regression or mediational analyses).

of the differences, whereas in many situations this may well be inaccurate. To the credit of Reicher and Wheeler, in addition to reporting inferential statistics related to the differences between means, they did also examine the number of people that showed the effect compared to those that did not. Indeed, in their experiments, the proportion was far higher than in ours. The point remains, however, that without investigating the data beyond the means, one's confidence that the means reflect the overall results should be low (e.g., Balota and Yap, 2011).

One heavy-handed solution to this problem would be to confine our theorizing to situations where the differences between conditions are so clear that there is no need for inferential statistics to determine whether or not differences are significant. An example of such a clear difference between conditions would be where 80% of participants in one condition show results that are higher/larger/better/faster than 80% of participants in another group. This would be a difference in performance that would be obvious, has a good chance of being replicated, and everyone would believe. Confining ourselves to effects that are this obvious would limit the number of phenomena that require explanation, and may reduce the current preponderance of seemingly unrelated phenomena and theories[5]. Although the shortcomings of inferential statistics have been discussed at great length elsewhere, and for some time now (Cohen, 1990, 1994; Hammond, 1996), it would appear that the message is not getting through. Indeed, when we asked some of our colleagues to read early drafts of this paper, a common response was along the lines "everyone knows this stuff." And yet we see little evidence of a change in behavior. An illustrative survey of the analysis methods used in research reported in 2012 in two prominent cognitive psychology journals is presented in **Table 2**. Perhaps the "everyone knows this stuff" response is a form of the hindsight bias (Hawkins and Hastie, 1990).

An illustration of how problematic averaged data can be comes from Heathcote et al. (2000). The target of their investigation was

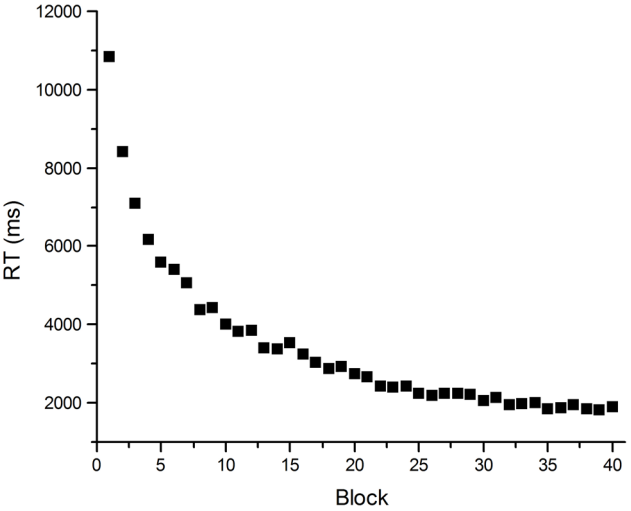


FIGURE 2 | Mean RT data from practice on an arithmetic task. Participants ($N = 40$) solved an equation [$(x^2 - y)/2 = A$) with eight repeating (x,y) pairs, determining whether A was "odd" or "even" for each pair, for 40 blocks of eight trials.

the Power Law of Learning. This refers to the observation that improvement in the speed of performing a task with practice has a characteristic pattern: performance improves by large amounts early in practice but these increments in performance get smaller as practice proceeds. The smooth trend in these learning curves can more often than not be described well by a power function (see **Figure 2**). Such curves have been observed in fields as disparate as cigar rolling (Crossman, 1959), reading mirror-reversed text (Kolers, 1976) and implicit memory (Kirsner and Speelman, 1996), and are similar to retention and forgetting curves in memory (Ebbinghaus, 1885). So ubiquitous is this observation that it has been said to comprise one of the few laws in psychology (Newell and Rosenbloom, 1981), and the one fact that requires explanation by any credible theory of skill acquisition (Anderson, 1982; Logan, 1988). Heathcote et al. (2000) however, called into question the lawfulness of this relationship between performance speed and practice. They demonstrated that power functions result from averaging any group data with a downward trend. The important point here is that power functions can appear in averaged data, even when they do not occur in individual data. Certainly if individual data is inspected, smooth learning curves are rarely observed. Although performance usually gets faster with practice on a task, performance from trial to trial almost never follows a smooth downward trend (see **Figure 3**).

So, why should theories attempt to explain power function learning if it does not actually exist in individual performance? The assumption of many cognitive theorists appears to be that individual performance does not reflect the real behavior they wish to explain, and that averaging is required over many trials and many people to remove the noise from the data in order for the real pattern to emerge. In the case of the Power Law of Learning, it appears that skill acquisition theorists assume that learning is smooth, and follows a power function, and therefore their

[5]Other solutions have been proposed before. One is to report effect sizes along with the results of significance testing (Cohen, 1994). This does not, however, address the issues we have identified because effect sizes are mostly used in comparisons of means. In this context the effect size is just a way of characterizing how large the difference is between the means, in relation to the variance observed. Another solution is to report confidence intervals with means (Cumming and Finch, 2005). Again this does not entirely solve the problem because a confidence interval is just a measure of average variance, and so glosses over details of a distribution of scores. Balota and Yap (2011) demonstrate that several parameters describing RT distributions can be psychologically meaningful.

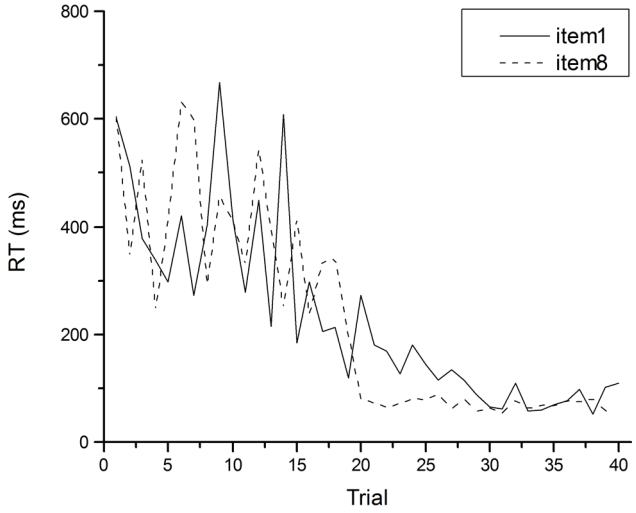


FIGURE 3 | An individual's RT as a function of practice for specific items in the arithmetic task of Figure 2.

theories must posit a learning mechanism that produces power function learning curves. It is possible, though, that this assumption is misguided. Rather than assuming that learning must follow a smooth trajectory, and so average data must be used to observe this smoothness, why not accept that the noise in data is an accurate reflection of the cognitive processes underlying performance? One theory of skill acquisition (Speelman and Kirsner, 2005) does take this position and considers noise in the data as the outcome of competition between cognitive processes striving to control performance. The lesson for all theories of cognitive processes, then, is that proposing mechanisms to explain mean performance may provide explanations of behavior that does not exist.

ILLUSTRATING THE PROBLEM IN BRAIN IMAGING RESEARCH

Psychology has recently faced a number of controversies that have caused us to take stock of our assumptions and practices as a discipline (e.g., Ritchie et al., 2012 and Roediger, 2012 on replication; Simonsohn, 2012 and Vogel, 2011 on data fraud). Little about the implications of these controversies is really new (Bakan, 1966; Meehl, 1978; Rosenthal, 1979; Cohen, 1994; Kirk, 1996; Rosenthal, 1966) though because of the relatively recent development and excitement of brain imaging research scrutiny in that domain is a particularly burgeoning field (Uttal, 2001; Vul et al., 2009; Carp, 2012), where sources of apparent stability or reliability in a behavior, effect or cognitive process are being increasingly questioned. While there are a number of different statistical and methodological bases for the various concerns raised by critics, the assumption of stability of function and consistency of operation across individuals that averages suggest so strongly is one that has not been carefully considered in much of the localization of function neuroscience literature.

Brain imaging research tends to produce colorful pictures of the brain with specific areas highlighted by bright colors to signify areas of high neural activity, typically associated with cognitive functioning of a particular type. Such pictures, however, are only generated through a process of combining activity patterns across many trials and many people. Data collected from an MRI machine are very noisy. Or at least, that is one interpretation. Another interpretation is that MRI machines produce an incredible amount of data. At any one moment in time the activity of neurons across the whole brain is inferred from the measurement of blood flow. If someone looking at a picture of the activity pattern hopes to see something easily interpretable, it is not surprising that the initial impression is that of a noisy and possibly random assortment of activations of varying degrees. However, if one assumes that hidden amongst the noise are areas of high activation where specific forms of cognitive processing are occurring, then one might consider looking for such areas by trying to eliminate the noise. Much of the noise comes from systematic sources such as the MRI machine and participant movement and so can be easily compensated for. Other noise, however, is seemingly random neural noise. Despite the impression provided by functional magnetic resonance imaging (fMRI) pictures, when areas of the brain are highly active, the other areas of the brain are not quiet. Determining the signal from the noise, then, becomes an important consideration when analyzing the activation patterns. Several methods are used to "clean up" the signal. One involves exposing individuals to many trials involving the same stimuli and requiring the same responses. The activation patterns from similar trials are then combined through averaging. This method assumes that the activation associated with particular stimuli and responses is similar each time. Another clean up method involves combining the average activation patterns from several individuals. One of the problems associated with this averaging step is that there are considerable individual differences in skull proportion. To combine patterns from different heads requires mathematically correcting each skull so that it matches the dimensions of a standard skull. This then ensures that activation patterns are combined from corresponding brain areas.

All of the averaging and correction involved in the analysis of fMRI activation patterns is concerning given our arguments about the mean. In particular we are concerned that fMRI researchers have designed their analytical tools to match their assumptions regarding what they will find in the data. Unfortunately, insufficient critical attention to these assumptions could mean that alternate hypotheses are ignored. For instance, when researchers combine fMRI activation patterns from many people, they assume that brain structures are similar across people, and they are responsible for similar functions in all people. Although we have no argument with the proposition that human brains share the same gross anatomy, we wonder about the assumption that specific localized areas within the cortex are responsible for specific cognitive functions. Some areas of the cortex have undeniable links with certain functions (e.g., the occipital lobe plays a major role in vision), but within certain areas, researchers often try to make the distinctions between an area that could be responsible for cognitive function X, whereas function Y is controlled by a different area (examples abound, see Gläscher et al., 2012 for a recent one). Further, we will get a stronger demonstration of this if we combine the activation patterns of many people, the more the better for the sake of statistical power. The problem with this strategy is that alternative hypotheses – that function is not localized, or that it is,

but is localized differently for each person – are ruled out by the methodology. The activation patterns that result from averaging over many trials and many people may not actually reflect what goes on in any one person's head. It may only be a reflection of what we would see if we undertake a lot of averaging. In other words, the activation pattern may not really exist except as some epiphenomenon of the methodology. This then raises a further problem – how can we make sense of group data when considering the case of an individual. Are we able to generalize a picture that has been derived from many data points to one data point? There are clear practical implications in cases involving brain damage and surgery.

Brain imaging research has over the past decade begun to move away from simple localization research, and significant developments have occurred in areas such as single-trial analyses (for a useful and brief map of that literature see Pernet et al., 2011). Indeed a recent study (Dinstein et al., 2012) that compared brain images from autistic adults and control subjects demonstrates how averaging brain images across trials and individuals can reveal a story quite different to the one that emerges from a focus on individual trial data. Others (Zilles and Amunts, 2013) have explicitly suggested that variability between subjects is not noise but important information. Reviewing neuroimaging work by Mueller et al. (2013) Zilles and Amunts examine a range of ways in which group-based analyses of neuronal structure can lead us to overlook information about individual differences in neuronal structure and change in structure over time that offers crucial clues to the processes underlying brain development. In particular, the use of group means allows different levels of individual variability to affect the sensitivity of methods used to find differences in brain regions (low variable regions will show small effects more noticeably).

While the logic of averaging is clearly problematic in the case of brain imaging research, and is under current active scrutiny, this issue nevertheless remains problematic for other areas of psychology too.

DOES THE MEAN UNCOVER OR IMPOSE UNIVERSALITY? REFLECTING FINDINGS IN PSYCHOLOGY UPON OUR OWN PRACTICES

The mean, like any piece of technology, is a tool. In itself it is impassive. Any use of a tool, however, is conducted on the basis of standards within a community of practice. While the mean itself is not laden with any particular theoretical assumptions our use of the mean is, and these assumptions are not without consequence.

Our purpose in the present paper is to encourage researchers to more frequently reflect on the fact that in focusing on the mean, in following a tendency to collapse things and encapsulate things into averages, we filter out individual differences and impose universality rather than finding it.

Molden and Dweck (2006) explore some of the ways in which a person's understanding or interpretation of a situation or phenomenon can have a dramatic effect on their behavior. Whether a person takes intelligence as a fixed capacity (what they term an "entity view"), for instance, or as something that can change or develop over time (an "incremental view") impacts on how a person performs in learning situations and how they respond to challenges or feedback. It would appear that individual differences in the meaningfulness of the situation can, sometimes dramatically, influence what a person does or what they are capable of, undermining any easy predictions based on what we might understand to be the "typical" cognitive system underlying such performance.

Summarizing several strands of such work, Molden and Dweck (2006) argue that while it is important for psychology to search for universals in behavior and cognition, these universals should be carefully described at the right level of abstraction. Our descriptions of the human psyche, when done in general or universal terms, potentially obscures the ways in which cognition, attitudes, values and behavior vary between people and between contexts.

This potentially limits our science in two important ways[6]. Firstly, by obscuring variation it biases our perceptions of the phenomena we study, making them appear more stable and determinate than may well be the case. Ironically, research in social psychology has warned us of such biases in human perception for decades. It is termed the fundamental attribution error (Jones and Harris, 1967; Ross, 1977), or correspondence bias, and is a notoriously difficult habit to break. Put simply, when we see another person act in a particular way we tend to see the behavior as dispositionally driven, rather than context-dependent. We are more likely to view the behavior as a stable characteristic of the individual rather than as a response to the specific vagaries of the circumstances in which the behavior occurs. To criticize much psychological research as falling prey to the fundamental attribution error would of course be glib and inaccurate but as a notion that is both provocative and evocative it is a useful tool with which to illustrate the problems of overlooking or downplaying variation in people's behavior or cognitive activity and summarizing outcomes with means alone. We rightly take pride in our use of objective tools in the conduct of our research and analysis but our exuberance for method can lead us to overlook the embedding of these tools in less objective assumptions and standards of practice that need periodic review. It is easy for us to make claims such as "the data show. . ." when the data can of course be used to support a number of different possible stories, once we have tamed it with data-cleaning techniques and stabilized the outcome with a single summary figure – the average.

The second way in which our science is limited by an overdependence on the mean as summary is in the generalization of results. The frequent use of the average as the sole description of a group's performance on a task, or measurement on a trait, characteristic or outcome, greatly limits our understanding of individual cases. It is well known that we cannot predict the individual case from statistics but where our discussions of measurements are presented almost exclusively in terms of averages we constrain ourselves to describing and discussing groups alone. It may be argued, reasonably, that decades of measurements have

[6]Molden and Dweck and other's research noted here uses mean group differences in precisely the way we are criticizing. The point is not that such research is not to be done, or is somehow valueless (far from it) but it must be interpreted and used judiciously, and we should look to what is possible given such conclusions rather than what is somehow essential. In the present case, we should be conservative about our use of certain statistical tools precisely because it is quite *possible* that it will negatively affect our science, as discussed further in the text.

shown that human beings are so variable in their responses as to make confident predictions of individual's actions to be foolhardy. There are so many variables, often interacting in non-linear ways, that generalization to the individual simply cannot be a reasonable aim of the discipline.

It is certainly the case that researchers routinely report variability measures (e.g., standard deviations, standard errors, confidence intervals) along with means. Despite this widespread reporting, however, one can question whether researchers are utilizing this information to temper their conclusions that are based upon the means. These measures generally provide information about the size of an underlying distribution but little information about its shape. Given the overwhelming focus upon mean scores, we wonder if the reporting of variability measures is merely an afterthought, or just fulfilling an expectation of journal reviewers and editors. More obviously, the reporting of sample variability measures completely overlooks the variability inherent in individual participants' responses – information regarding this variability is eliminated by using subject means to reflect each person's performance.

Nonetheless, examining ranges of scores and variability as phenomena of interest in and of themselves would provide us with a context within which to frame individual observations, the better to understand what the possibilities are such that we can then make a more informed decision about what the *probabilities* might be in the individual case. Rather than seeing outliers as unclean, aberrations or errors that should be excised before the real work begins, they provide us with information on what is possible. Rather than trying to prophecy single specific outcomes, which would likely be unsuccessful, a describing of the landscape of possibilities would provide useful insight in many behavioral contexts.

Context effects are ubiquitous in all areas of psychological research. Our habit of describing things in terms of means rather than ranges and distributions tends to reduce analysis of this fact into a list of independent observations, a shopping list of possibilities with little to relate the differences in cognitive function from one situation to the next. The very term "context" is frustratingly difficult to define, and varies in use from experiment to experiment, researcher to researcher, a lack of discipline that isolates the work of different individuals and thus obscures what relationships exist between the various independent observations, making it difficult if not impossible to overcome the "shopping list" state of our current understanding. (It is left as an exercise for the reader to review what the "context effects" are in their own domain of interest and to examine just what is being considered as "context," and the criteria on which that decision is based.)

We would argue that the unguarded use of the mean to summarize outcomes from different experiments suppresses the perception of both variability and continuity between results, tempting us to see the differences as more stable and certain than they really are, and leading to the balkanisation of research that limits our insights into psychological functioning. In this, we suggest that the mode of analysis and description inherent in the use of the mean as principal summary statistic is very similar to that of the cognitive linguistic phenomenon of nominalization as described by Barsalou et al. (2010), p. 350,

> ...it is possible to conceptualize nouns in decontextualized ways, and these decontextualizations play important roles. We err, however, when we mistakenly believe that these decontextualized mechanisms refer to meaningful entities in isolation, and forget that they operate intrinsically in contexts and depend on contexts for their realization. The mechanism indexed by a noun integrates a large system of situated patterns, with this system usually producing an emergent form well-suited to the current situation.

Similarly, the behavior or cognitive activity indexed by a mean of measured performance is a collection of context-sensitive processes that likely include much more than the specific independent variable with which that mean is explicitly associated in a given study. Though it is inherent in psychological training that we be critical and circumspect in our assessment of reported results, we are not immune to the biases that we report in our participants' behavior.

In no sense do we suggest that the mean is somehow wrong. The problem is rather that it is so satisfying. Decades of research on attribution biases and Barsalou et al.'s (2010) work on nominalization suggests that the kind of encapsulated and stable idea of performance that the mean suggests is an enticing, seductive view (at least for the Western majority involved in high "impact" psychological research). The basic aims of research, insofar as it entails a search for the general and the universal fill our perceptions and interpretations of data and settle standards of practice that lean heavily toward the stable, reliable and consistent. We thus suggest that the problematically uncritical use of the mean is an expression of an unreflectively held view of the psychological system. Re-consideration of our statistical tools will also involve some re-consideration of our theoretical standpoint and the standard ways in which we formulate research questions.

Though there are certainly domains and approaches within psychology that emphasize contextualized performance and situational variability (see for instance Barrett et al., 2010 for a survey of recent cognitive work; see also the much discussed situational view of personality by Mischel and Shoda, 1995; Mischel, 2004), the search for general capacities and universal functions is by far the more common.

SUPPRESSING THE ASSUMPTION OF STABILITY

There are alternatives to thinking about psychological mechanisms as shared and stable characteristics of the human species that do not lead us into despair or pessimism about the possibility of a unified and systematic theory of psychology. The development, over the past two decades, of modes of thinking that place great emphasis on individual developmental dynamics, the (often messy) details of a cognitive agent's actual, real, bodily interaction with its environment, provide us with an approach that allows for more nuanced, dynamic perspectives on psychology and psychological processes.

There are a number of these different ways of thinking. They are not necessarily commensurable with one another and as yet do not offer a single coherent vision of psychology that might be recognizable as a "paradigm" in the Kuhnian sense, to which we could leap in some revolutionary fervor. However, what these different approaches make clear, having been developing around the fringes of the discipline for decades and gradually encroaching further

into mainstream research, is that valuable, fruitful research can be done in which the emphasis is placed on the dynamics of change in psychological processes over time, and in which the complexity of interactions between the individual characteristics of both the person and the environment in which they are acting can be accounted for and incorporated into scientific psychological theory. Such approaches do not suppress variation in the behaviors of people but rather see it as a rich resource for understanding how psychology interacts with context. Similarly, while these approaches use a variety of statistical tools other than the standard significance testing that remains the mainstream, the arithmetic mean is still used, but its use does not require the assumptions of underlying "true" values clothed in noise with which we take issue in this paper.

The developmental dynamics of Thelen and Smith (1994) are a perfect example of this focus on processes and change over rigid structure. Thelen and Smith examine development as a contextualized process of interplay between the child and their environment, providing evidence for the growth of motor skills not as the blossoming of standard, universal cognitive capacities (true values to be approximated with averaged observations) but as the coping of the individual child with the demands of their idiosyncratic histories. The differences between children in their development has at least as much to tell us about how development occurs than the similarities.

The idea that the cognitive system is not rigidly specified, but is in fact supple and responsive (over a number of timescales) to the quirks, specifics and details of the environment in which it operates is summed up by Clark's (1997) description of the "soft assembly" of cognitive function. Rigidly or "hard assembled" systems have a fixed structure and mode of operation. There is a "right" way to describe how the different components of the system relate to one another, an ideal of the system that is, in some fundamental way, correct. Not so for soft assembled systems.

Soft assembled systems tend to have loosely inter-connected components, less fixed positions within a structure so much as a pool of potential resources that can be organized within various constraints in response to situations and task demands. There is no ideal of how such a system should be organized, no schematic that can be drawn that captures the correct way in which the components might relate to one another, as these things will vary continuously depending on contexts, individual histories and immediate requirements. Soft assembled systems tend not to use central controllers but rather they self-organize, with task-specific activity emerging from the dynamic interaction between components and environment. This can happen either over quite brief timescales, or more slowly over longer periods.

The net result, if cognitive activity is assembled in such a manner, is that similar behavior might in fact be the result of quite differently organized psychological processes. There is no "correct" mapping of the psychological system, no signal about cognitive structure being hidden by the noise of individual variation. Such a theoretical standpoint eschews assumptions of single true values to be sought in the noise of individual variation and measurement error. What must be understood is the dynamics of response to situations over time, with an appreciation that different individual histories will often result in quite differently arranged but similarly performing psychological systems. What is more, it may be the case that even within a single individual over particular timescales (those associated with learning in its many forms) we might see the structure and functioning of the cognitive system changing dramatically.

UNDERSTANDING SIMILARITY AND STABILITY IN BEHAVIORAL PERFORMANCE

The idea of a stable and shared set of basic cognitive processes underlying some of our use of the mean is not entirely an assumption. It is rooted in the success of our everyday interactions, the ease with which we can coordinate with one another, share experiences and activities. A critical reader would no doubt at this point be arguing that the statistical tests we typically use in data analysis invariably take the variance or deviations within the sample into account, while also emphasizing the plain fact that while it might be true that everyone is unique, it is plainly true that we share a great deal. People vary, sure, but looking around, they do not vary nearly as much as they could in most cases, and most of what differences do exist seem to be quite subtle – certainly nothing requiring any fundamental re-think of our use of statistics or theoretical perspective.

Amongst these new approaches to psychological research, how is this simple truth about the similarity of human beings to be captured and explained?

There are certainly some things that human beings generally share (though there are very few that are genuinely universal). Our basic body plan, our nutritional requirements, the range of physical stimuli to which we are sensitive and with which we can interact, these tend to vary within fairly narrow ranges. These shared constraints on our behavior will provide particular channels for developmental change, channels that will be structured further by the cultural provision of particular developmental tasks and demands. Each human being follows a unique developmental trajectory but there are constraints on that trajectory. A key observation here, however, is that this more developmental mode of explanation suggests that the reason for people's similarities is not the inexorable unfolding of a pre-specified and consistently presenting cognitive system, but the shared constraints on development, which may specify the ends (consistency in behavior), but will typically underspecify the means.

The developmental work of Thelen and Smith (1994) once again offers us some examples. Their much cited work examining the development of reaching and grasping in two infants, Hannah and Gabriel, describes how features of the children's bodies, their intrinsic dynamics, idiosyncrasies of energetics or even simply mass, mean that each child has a different developmental task in order to achieve the same outcome. Whereas the excitable Gabriel must learn to draw energy out of the whole reaching system and slow his movements down if he is to manage to get his hand successfully to a target object, the more placid Hannah must learn to put more oomph and effort in to achieve the same result. Karmiloff-Smith (2000) describes similar differences, this time not in order to explain individual differences but rather to explain the response to different developmental constraints in the performance of the "standard" function of face recognition for people with or without Williams syndrome (WS). She argues that similarly proficient behavior of people with or without WS in the

overall task of face recognition is underpinned by quite variant collections of more basic cognitive skills (e.g., recognizing facial identity, facial emotion, eye-gaze direction, or lip-reading). That is, the same behavioral outcome can be the result of quite different forms of underlying process. This is a clear case where one researcher's noise is another researcher's signal.

Typically as researchers it is precisely those choices and manipulations of the environment that will ensure the highest probability of similar performance are those most highly valued in experimental settings. In most of our research we take great pains to limit the range of participants' behaviors in order to make consistency and similarity the most likely outcome. Gross variability, or indeed any variability not directly attributable to the chosen independent variable is considered a sign of a poorly designed study – other potential sources of change and difference are suppressed. There are certainly times when this is desirable. We suggest, however, that such practices have been adopted as standard and implemented by many researchers without due consideration of their valid range of application. Recently developing theoretical perspectives within the dynamical family of viewpoints push explicit reflection on these questions to the fore once again. We consider this a very positive development, one which will not suppress the use of the mean in our research, but will hopefully suppress its use in an uncritical, or overly focused manner.

CONCLUSION

Our point in raising these issues is not to suggest that psychology wean itself off use of the mean, or to go cold turkey on averages. Such inane recommendations would deserve the disdain with which they would inevitably be met. However, we do argue for a more careful, critical and explicit use of averages in the discussion of measurements and the reporting of results. Specifically, we argue that the mean must not be used without reflection on the theoretical assumptions and frameworks that are underlying its use and we suggest that in the typical case a theoretical perspective closer to that of dynamical systems will be more appropriate, providing more context and a fuller picture of the behavior in question from the data observed.

The average provides us with important and useful information but we see its use in summarizing and analyzing groups to suppress important individual differences in behavioral and cognitive performance as having become unbalanced. The range and variance of scores in distributions should be reported as frequently and clearly as averages and should temper our easy acceptance of the mean as representative of the numerous individual people whose behavior or characteristics are being recorded. Tukey (1977) pioneered graphical techniques for presenting such information. An excellent recent example is provided by Doherty et al. (2013), whose Figure 4 presents the means from a one-way ANOVA design, along with all of the data that featured in the analysis. This figure not only depicts the relationship between the independent variable and means on the dependent variable, but it also reveals the extent to which the relationship exists amongst the individual observations, and represents the overlap between conditions more completely than a group of error bars or confidence intervals. Other examples already mentioned are those of Marmolejo-Ramos and Matsunaga's (2009) work on graphical techniques in exploratory data analysis and Balota and Yap's (2011) suggestions about "moving beyond the mean" in analysis of reaction time curves. Readers are also pointed toward Landau's (2002) introduction to survival curves, which allow for the mapping of relationships between variables and outcomes over time in a simple but clear manner.

Rather than focusing purely on the question of whether a difference exists, our aim should be to use statistics to illustrate and characterize the range of measurements recorded as fully as possible. By using the range of quantitative options available more fully (range, median, variance, and others) we can provide a better qualitative appreciation of the behaviors we observe, a richer and more nuanced picture of the phenomena that we are interested in describing, explaining and predicting. This will also allow our predictions to become much more interesting – not just whether one group will be bigger or faster or more but what the range or distribution of outcomes are likely to be depending on the size of the sample or its composition. Further, we could examine whether there are differences on a range of variables (e.g., working memory capacity, IQ, reading speed) between people who do and do not show the average target effect. This will provide us with a richer data source that may reveal more about why some people exhibit the effect and others do not, and this would expose more information about the mechanism under scrutiny.

We would also do well to expand the set of tools available to us. The history of psychology is rich in alternate methods of analyzing behavior that do not rely on averaging group data. Research in psychophysics regularly analyses data from individual subjects; neuropsychology has a long history of single case studies; Piaget's theories were developed on the basis of analyses of the behavior of a few subjects. In cognitive psychology, however, although there are many researchers who fit mathematical models to individual data (Lewandowsky and Farrell, 2011), the modal behavior is to focus on grouped data and average performance (see **Table 2**). Lewandowsky and Farrell, 2011 (p. 106) suggest

> that it may be advisable to fit one's models to both individual data and aggregate data; if both yield the same result, one can have increased confidence in the outcome. If they diverge, one ought to rely on the fit to individuals rather than the aggregate.

With such data, dynamical and complex systems thinking offers rich possibilities for alternate modes of investigation, as does Bayesian analysis. Of course, these new instruments would not exempt us from our role as sensitive, judicious and critical tool users any more than would our more widely practiced and familiar analytical techniques.

The mean's many roles should be clear in our minds as we design and conduct our experiments, as we take measurements, and carry out analyses. Interpretations of the results should be limited accordingly. Such critical consideration of the mean may prompt us to broaden our methodological horizons, balancing a sensitivity to the potentially universal and broadly shared with the unique, the variable and the idiosyncratic. Ultimately, we should be mindful of the purposes for which we are using the mean and more importantly, the things we are trying to reflect by using the mean. We should consider the degree to which we can assume that the people from whom we take measurements and calculate

average scores all possess a similar cognitive mechanism that underlies the performance we have measured. If we think there is likely to be a high degree of similarity in mechanisms, then reflecting that performance with a mean is justified. Otherwise, the mean will severely obscure variations in performance and hence the variety of cognitive mechanisms possessed by people.

Finally, as researchers, we should deliberate over whether we are making the assumptions about the mean that have been highlighted here. If these assumptions are implicit in our methodology, then we should consider whether each is justifiable in our particular research context. If they are not justifiable, then alternate tools may be necessary.

REFERENCES

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychol. Rev.* 89, 369–406. doi: 10.1037/0033-295X.89.4.369

Bakan, D. (1966). The test of significance in psychological research. *Psychol. Bull.* 66, 1–29. doi: 10.1037/h0020412

Balota, D. A., and Yap, M. J. (2011). Moving beyond the mean in studies of mental chronometry the power of response time distributional analyses. *Curr. Dir. Psychol. Sci.* 20, 160–166. doi: 10.1177/0963721411408885

Barrett, L. F., Mesquita, B., and Smith, E. R. (2010). "The context principle," in *The Mind in Context*, 1st Edn, eds B. Mesquita, L. F. Barrett, and E. R. Smith (London: Guilford Press).

Barsalou, L., Wilson, C. D., and Hasenkamp, W. (2010). "On the vices of nominalization and the virtues of contextualizing," in *The Mind in Context*, 1st Edn, eds. B. Mesquita, L. F. Barrett, and E. R. Smith (London: Guilford Press).

Boring, E. G. (1920). The logic of the normal law of error in mental measurement. *Am. J. Psychol.* 31, 1–33. doi: 10.2307/1413989

Carp, J. (2012). The secret lives of experiments: methods reporting in the fMRI literature. *Neuroimage* 63, 289–300. doi: 10.1016/j.neuroimage.2012.07.004

Clark, A. (1997). *Being There: Reuniting Brain, Body and World*. Boston: MIT Press.

Cohen, J. (1977). *Statistical Power Analysis for the Behavioural Sciences*. New York: Academic Press.

Cohen, J. (1990). Things I have learned so far. *Am. Psychol.* 45, 1304–1312. doi: 10.1037/0003-066X.45.12.1304

Cohen, J. (1994). The earth is round ($p < .05$). *Am. Psychol.* 49, 997–1003. doi: 10.1037/0003-066X.49.12.997

Crossman, E. R. (1959). A theory of the acquisition of speed-skill. *Ergonomics* 2, 153–166. doi: 10.1080/00140135908930419

Cumming, G. (2012). *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. New York: Routledge.

Cumming, G., and Finch, S. (2005). Inference by eye: confidence intervals and how to read pictures of data. *Am. Psychol.* 60, 170–180. doi: 10.1037/0003-066X.60.2.170

Dancey, C. P., and Reidy, J. (2008). *Statistics Without Maths for Psychology: Using Spss for Windows*. London: Pearson Education.

Dinstein, I., Heeger, D. J., Lorenzi, L., Minshew, N. J., Malach, R., and Behrmann, M. (2012). Unreliable evoked responses in Autism. *Neuron* 75, 981–991. doi: 10.1016/j.neuron.2012.07.026

Doherty, M. E., Shemberg, K. M., Anderson, R. B., and Tweney, R. D. (2013). Exploring unexplained variation. *Theory Psychol.* 23, 81. doi: 10.1177/0959354312445653

Ebbinghaus, H. (1885). *Memory*. New York: Columbia University/Dover.

Gläscher, J., Adolphs, R., Damasio, H., Bechara, A., Rudrauf, D., Calamia, M., et al. (2012). Lesion mapping of cognitive control and value-based decision making in the prefrontal cortex. *Proc. Natl. Acad. Sci. U.S.A.* 109, 14681–14686. doi: 10.1073/pnas.1206608109

Halloy, S. R. P. (1998). A theoretical framework for abundance distributions in complex systems. *Complex. Int.* 6, 12.

Hammond, G. (1996). The objections to null hypothesis testing as a means of analysing psychological data. *Aus. J. Psychol.* 48, 104–106. doi: 10.1080/00049539608259513

Haslam, S. A., and McGarty, C. (2003). *Research Methods and Statistics in Psychology. Sage Foundations of Psychology Series*, 2nd Edn. London: SAGE.

Hawkins, S. A., and Hastie, R. (1990). Hindsight: biased judgments of past events after the outcomes are known. *Psychol. Bull.* 107, 311–327. doi: 10.1037/0033-2909.107.3.311

Heathcote, A. S., Brown, S., and Mewhort, D. J. K. (2000). The power law repealed: the case for an exponential law of practice. *Psychol. Bull. Rev.* 7, 185–207. doi: 10.3758/BF03212979

Heathcote, A., and Love, J. (2012). Linear deterministic accumulator models of simple choice. *Front. Psychol.* 3:292. doi: 10.3389/fpsyg.2012.00292

Holden, J. G., Van Orden, G. C., and Turvey, M. T. (2009). Dispersion of response times reveals cognitive dynamics. *Psychol. Rev.* 116, 318–342. doi: 10.1037/a0014849

Howell, D. C. (2002). *Statistical Methods for Psychology*, 5th Edn. Pacific Grove, CA: Duxbury.

Howitt, D., and Cramer, D. (2011). *An Introduction to Statistics in Psychology*, 5th Edn. Harlow: Prentice Hall.

Jones, E. E., and Harris, V. A. (1967). The attribution of attitudes. *J. Exp. Soc. Psychol.* 3, 1–24. doi: 10.1016/022-1031(67)90034-0

Karmiloff-Smith, A. (2000). "Why babies' minds aren't Swiss Army Knives," in *Alas, Poor Darwin*, eds H. Rose and S. Rose (London: Jonathan Cape). 144–156.

Kirk, R. E. (1996). Practical Significance: a concept whose time has come. *Educ. Psychol. Meas.* 56, 746–759. doi: 10.1177/0013164496056005002

Kirsner, K., and Speelman, C. (1996). Skill acquisition and repetition priming: one principle, many processes. *J. Exp. Psychol. Learn. Mem. Cogn.* 22, 563–575. doi: 10.1037/0278-7393.22.3.563

Kolers, P. A. (1976). Reading a year later. *J. Exp. Psychol. Hum. Learn. Mem.* 2, 554–565. doi: 10.1037/0278-7393.2.5.554

Landau, S. (2002). Using survival analysis in psychology. *Unders. Stat.* 1, 233–270. doi: 10.1207/S15328031US0104_03

Lewandowsky, S., and Farrell, S. (2011). *Computational Modeling in Cognition: Principles and Practice*. Thousand Oaks, CA: Sage Publications.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychol. Rev.* 95, 492–527. doi: 10.1037/0033-295X.95.4.492

Marmolejo-Ramos, F. and Matsunaga, M. (2009). Getting the most from your curves: exploring and reporting data using informative graphical techniques. *Tutor. Quant. Methods Psychol.* 5, 40–50.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *J. Consult. Clin. Psychol.* 46, 806. doi: 10.1037/0022-006X.46.4.806

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychol. Bull.* 105, 156–166. doi: 10.1037/0033-2909.105.1.156

Mischel, W. (2004). Toward an integrative science of the person. *Annu. Rev. Psychol.* 55, 1–22. doi: 10.1146/annurev.psych.55.042902.130709

Mischel, W., and Shoda, Y. (1995). A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychol. Rev.* 102, 246–268. doi: 10.1037/0033-295X.102.2.246

Molden, D. C., and Dweck, C. S. (2006). Finding "Meaning" in psychology: a lay theories approach to self-regulation, social perception, and social development. *Am. Psychol.* 61, 192–203. doi: 10.1037/0003-066X.61.3.192

Mueller, S., Wang, D., Fox, M. D., Yeo, B. T. T., Sepulcre, J., Sabuncu, M. R., et al. (2013). Individual variability in functional connectivity architecture of the human brain. *Neuron* 77, 586–595. doi: 10.1016/j.neuron.2012.12.028

Newell, A., and Rosenbloom, P. S. (1981). "Mechanisms of skills acquisition and the law of practice," in *Cognitive Skills and their Acquisition*, ed. J. R. Anderson (Hillsdale, NJ: Erlbaum).

Novick, M. R. (1966). The axioms and principal results of classical test theory. *J. Math. Psychol.* 3, 1–18. doi: 10.1016/0022-2496(66)90002-2

Patel, R., Spreng, R. N., and Turner, G. R. (2013). Functional brain changes following cognitive and motor skills training: a quantitative meta-analysis. *Neurorehabil. Neural Repair* 27, 187–199. doi: 10.1177/1545968312461718

Pernet, C. R., Sajda, P., and Rousselet, G. A. (2011). Single-trial analyses: why bother? *Front. Psychol.* 2:322. doi: 10.3389/fpsyg.2011.003322

Ratcliff, R., and McKoon, G. (2008). The diffusion decision model: theory and data for two choice decision tasks. *Neural Comput.* 20, 873–922. doi: 10.1162/neco.2008.12-06-420

Reicher, G. M. (1969). Perceptual recognition as a function of meaningfulness of stimulus material. *J. Exp. Psychol.* 81, 275–280. doi: 10.1037/h0027768

Ritchie, S. J., Wiseman, R., and French, C. C. (2012). Replication, replication, replication. *Psychologist* 25, 346–348.

Roediger, H. L. (2012). Psychology's woes and a partial cure: the value of replication. *Observer* 25. Available at: http://www.psychologicalscience.org/index.php/publications/observer/2012/february-11-2012-observer-publications/psychology's-woes-and-a-partial-cure-the-value-of-replication.html

Rosenthal, R. (1966). *Experimenter Effects in Behavioral Research.* East Norwalk, CT: Appleton Century Crofts.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychol. Bull.* 86, 638. doi: 10.1037/0033-2909.86.3.638

Ross, L. (1977). "The intuitive psychologist and his shortcomings: distortions in the attribution process," in *Advances in Experimental Social Psychology*, Vol. 10, ed. L. Berkowitz (New York: Academic Press), 173–220.

Simonsohn, U. (2012). *Just Post It: The Lesson from Two Cases of Fabricated Data Detected by Statistics Alone.* Available at SSRN: http://ssrn.com/abstract=2114571

Speelman, C. P., and Kirsner, K. (2005) *Beyond the Learning Curve: The Construction of Mind.* Oxford: Oxford University Press.

Thelen, E., and Smith, L. B. (1994). *A Dynamic Systems Approach to the Development of Cognition and Action.* Cambridge, MA: MIT Press.

Tukey, J. W. (1977). *Exploratory Data Analysis.* Reading, MA: Addison-Wesley.

Uttal, W. R. (2001). *The New Phrenology: The Limits of Localizing Cognitive Processes in the Brain.* Cambridge, MA: MIT Press.

Vogel, G. (2011). *Report: Dutch "Lord of the Data" Forged Dozens of Studies. Science Insider.*

Vul, E., Harris, C., Winkielman, P., and Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspect. Psychol. Sci.* 4, 274–290. doi: 10.1111/j.1745-6924.2009.01125.x

Wheeler, D. D. (1970). Processes in word recognition. *Cogn. Psychol.* 1, 59–85. doi: 10.1016/0010-0285(70)90005-8

Zilles, K., and Amunts, K. (2013). Individual variability is not noise. *Trends Cogn. Sci.* 17, 153–155. doi: 10.1016/j.tics.2013.02.003

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 16 April 2013; accepted: 29 June 2013; published online: 23 July 2013.
*Citation: Speelman CP and McGann M (2013) How mean is the mean? Front. Psychol. **4**:451. doi: 10.3389/fpsyg.2013.00451*
This article was submitted to Frontiers in Quantitative Psychology and Measurement, a specialty of Frontiers in Psychology.
Copyright © 2013 Speelman and McGann. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in other forums, provided the original authors and source are credited and subject to any copyright notices concerning any third-party graphics etc.

frontiers in Psychology

EDITORIAL
published: 19 August 2016
doi: 10.3389/fpsyg.2016.01234

Editorial: Challenges to Mean-Based Analysis in Psychology: The Contrast Between Individual People and General Science

Craig P. Speelman[1]* *and Marek McGann*[2]

[1] School of Psychology and Social Science, Edith Cowan University, Joondalup, WA, Australia, [2] Department of Psychology, Mary Immaculate College, University of Limerick, Limerick, Ireland

Keywords: mean, average, variability, inference, groups, individuals

The Editorial on the Research Topic

Challenges to Mean-Based Analysis in Psychology: The Contrast Between Individual People and General Science

In a recent paper we (Speelman and McGann) argued that psychology's reliance on data analysis methods that are based on group averages has resulted in a science of group phenomena that may be misleading about the nature of and reasons for individual behavior. The paper highlighted a tension between a science in search of general laws on the one hand, and the individual, variable, and diverse nature of human behavior on the other. Two central traditions in psychology are challenged by this tension: (1) data is collected from a large number of people and distilled into a handful of parameters that reflect the middle of a distribution of scores and the average variation around that mid-point, and (2) theories are developed to explain the average performance of the group. The disjunction between group-based measurements and the actual psychology of individual people raises specific concerns in both research and applied professional domains of psychology. For instance, a clinician who reads in a report that Therapy A leads to a significantly greater improvement in depression than Therapy B might be tempted to adopt Therapy A in her practice. But what are the odds that Therapy A will be the best option for the next depressed client to walk in her door? What does an observation that, on average, people find it easier to identify letters presented on a screen when they are presented at the end of a word than when presented in isolation actually tell us about the specific cognitive processes occurring in specific people's activities? Are we justified in interpreting this result as reflecting something about the way every person's mind processes letters and words? To what extent should we explore the prevalence of this pattern of responding before we start making claims about cognitive mechanisms that are general to all humans?

We argued that more explicit and careful justifications are required for the common practice in psychology of extrapolating from average data to general laws, but also from general laws to explanations of individual behavior. Given the ability of humans to adapt to their environments, it would seem unlikely that everyone would develop identical cognitive processes for any given task. As a result, developing general theories about any given task, and using those theories to develop methods for clinical interventions or educational purposes would seem a risky endeavor.

This Research Topic explored this concern about the pitfalls of using the mean for the basis of psychological science. The problem is universal in its applicability to psychology, and opinion papers, reviews, and original empirical research from all areas of the discipline were invited.

A total of 16 authors contributed 9 articles to the Topic. The range of issues that the authors viewed through the lens provided is impressive. These articles follow two principal themes. The first concerns the relationship between theory and different statistical techniques, and how a more

OPEN ACCESS

Edited and reviewed by:
Jill L. Adelson,
University of Louisville, USA

***Correspondence:**
Craig P. Speelman
c.speelman@ecu.edu.au

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *18 June 2016*
Accepted: *03 August 2016*
Published: *19 August 2016*

Citation:
Speelman CP and McGann M (2016) Editorial: Challenges to Mean-Based Analysis in Psychology: The Contrast Between Individual People and General Science. Front. Psychol. 7:1234. doi: 10.3389/fpsyg.2016.01234

comprehensive understanding of psychology demands a more varied (and perhaps more precise) set of investigative techniques. The second theme concerns more fine-grained technicalities, and the papers here illustrate the practical significance of understanding the relationship between measures of central tendency and other characteristics of the data sets that give rise to them.

Papers in the first theme explore ways in which we can discipline our data collection to avoid the traps of logic associated with careless use of averages. Campitelli, for instance, argues that psychology typically produces imprecise theories and so tends to fit its research questions to the available statistical tools. He advocates for the development of more precise theories and describes four analytical methods that he has used to answer precise research questions and which do not require the calculation of the mean. Grice also recommends the development of theoretical models that are person-centered, rather than group based, and so do not require aggregate statistics, such as the mean, to evaluate. Such an approach is perhaps more akin to a detective gathering clues to solve a mystery, enabling investigators to gather information and test specific models based on patterns of collected evidence, rather than on the success or failure of individual observations.

McAuliffe and McGann explore one particular way to gather information about the context of behavioral measurements that may highlight variability, and enable an exploration of that variability within standard laboratory tasks. They suggest adapting Hurlburt's descriptive experience sampling method for the laboratory in order to enable interrogation of behavioral performance in terms of the details and variety of individual experiences reported by participants during a given task.

Finally in this group, Kirsner's article describes the long and convoluted process involved in predicting the locations of two related shipwrecks. He shows how aggregating many disparate pieces of information pointed to the most accurate locations for these wrecks, a process he likens to the calculation of the mean or population parameter, and so highlights a situation where multiple perspectives provide a kind of parallax that can be used to bring a single target into focus, rather than depending on multiple measurements of the same variable to average out noise.

Complementary to these explorations of alternative methodological or analytic approaches are papers that illustrate and explicate more specific technical problems with various uses of the mean. In each of these papers the relationship between the mean and other aspects of the data in question can have a substantial impact on the validity of our inferential techniques, and the kinds of conclusions we might draw.

Speelman and Muller Townsend examined the extent to which average group performance can mask the heterogeneity that exists between the members of a group. They demonstrated that a substantial proportion of participants do not demonstrate a transition from controlled to automatic performance in a standard training experiment, despite the fact that the group results suggested such a transition occurred.

Looking at linear mixed-effects models as a set of analytical methods for overcoming problems associated with the mean, Lo and Andrews examine their ability to satisfy normality assumptions without the need to resort to transformation allowing investigators to work much more closely to the raw data themselves.

Hamaker and Grasman demonstrate how decisions about the centering methods used in cluster analysis can affect the ultimate solution, and that this affects levels of a multi-level autoregressive model differently. Their work emphasizes once again the importance of careful, deliberate use of our analytical tools, and that effective statistics rely on clearly set out, and explicit theorizing. Schuurman et al. work complements this somewhat examining the effect of including multiple sources of variation into a model, specifically focusing on noise in data. Mostly associated with measurement error, they show that noise can have a substantial effect on parameter estimation in autoregressive modeling. On the basis of their simulation study, they conclude that incorporating this noise into an analysis results in more accurate estimation.

Finally, Trafimow discusses how the meaning that can be attributed to the value of a sample's standard deviation can depend on the value of the sample mean, and vice-versa. Using a newly defined "coefficient of centrality" (the reciprocal of the coefficient of variation) as a means of relating the mean and standard deviation, he recommends that researchers routinely consider standard deviations when interpreting means. While other papers perhaps illustrate more dramatic departures from currently widely used practices in psychological statistics, Trafimow's work shows how relatively modest changes in our approach can provide quite striking improvements in our understanding.

Psychology as a discipline has been facing challenges that are not simply statistically significant, but practically, and perhaps fundamentally so. In our 2013 paper we noted that much in our argument was not particularly novel to psychologists, but despite a background or low-level awareness of possible problems, as a profession we have rather stubbornly pushed on with an uncritical or unthinking use of averages in our descriptions of groups, and a suppression of variation in our interpretation of results. The papers in this collection include a range of perspectives that provide concrete examples of how to approach research design, data collection, and analysis differently. No one contribution will provide a solution to our multifarious challenges, but nor should it. Our subject matter is complex and subtle, our investigations and methodological techniques will need to be equally so.

AUTHOR CONTRIBUTIONS

All authors listed, have made substantial, direct and intellectual contribution to the work, and approved it for publication.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Speelman and McGann. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

OPINION
published: 08 September 2015
doi: 10.3389/fpsyg.2015.01379

Answering research questions without calculating the mean

Guillermo Campitelli *

School of Psychology and Social Science, Edith Cowan University, Perth, WA, Australia

Keywords: research methods, mean, variability, deliberate practice, expertise

In an important theoretical article Speelman and McGann (2013) indicated that psychological researchers tend to use statistical procedures that involve calculating the mean of a variable in an uncritical manner. A typical procedure in psychological research consists of calculating the mean of some dependent variable in two or more samples and to present those means as summaries of the samples. The next step is to use some statistical technique (e.g., *t*-test, ANOVA) in order to be able to determine the probability of finding the observed differences between means in those samples given that the difference between the means of the populations from which the samples were extracted is zero. If this probability is very low (i.e., <0.05) the psychological researcher decides that the difference between the means of the populations of interest is not zero.

This procedure—the null hypothesis statistical significance testing (NHST) procedure–has received a huge number of criticisms, which are beyond the scope of this article. However, I would like to present the anecdote told by Cohen (1994), not only to criticize the NHST procedure itself but also the uncritical manner in which the procedure is used. Cohen tells us that a colleague hypothesized that a rare disease did not exist in a population; he then collected a sample of 30 individuals and found that one of them had the disease. He then wondered what type of significance test should be used in this situation. Obviously, the existence of one case with the disease is enough evidence to refute the hypothesis, but the uncritical search for a hypothesis testing procedure precluded the researcher from seeing the obvious.

This anecdote nicely dovetails with Speelman and McGann's (2013) assertion that psychological researchers tend to use procedures that involve calculating means in an uncritical manner. The goal of this article is to emphasize that there are procedures that do not involve calculating means, which are perfectly sound to answer research questions. In the following sections I will present the endeavor that other colleagues in the field of psychology of expertise and I embarked on with the purpose of testing hypotheses of the deliberate practice framework (Ericsson et al., 1993). I will present four measures that did not involve calculating the mean (i.e., variability, a value, a case, and distributions) I have used in my research to answer research questions. Before that I briefly explain the deliberate practice framework.

Deliberate Practice Framework

Ericsson et al. (1993) presented the deliberate practice framework of expert performance. The framework provides recommendations of how to conduct research in the field of expertise, it defines what deliberate practice is and it states that abundant deliberate practice constitutes a *necessary* and a *sufficient* condition to achieve high levels of expertise (see Campitelli and Gobet, 2011; Ericsson, 2014; Hambrick et al., 2014a,b for a discussion about the hypotheses of the deliberate practice framework).

Ericsson et al. (1993) defined deliberate practice as engaging in highly structured domain-specific activities deliberately developed to correct technical mistakes and to improve performance, which are conducted with high concentration levels and are followed by

OPEN ACCESS

Edited by:
Marek McGann,
Mary Immaculate College, Ireland

Reviewed by:
Min Liu,
University of Hawaii at Manoa, USA
Brody Heritage,
Murdoch University, Australia

***Correspondence:**
Guillermo Campitelli,
gjcampitelli@gmail.com

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *11 June 2015*
Accepted: *27 August 2015*
Published: *08 September 2015*

Citation:
Campitelli G (2015) Answering research questions without calculating the mean. Front. Psychol. 6:1379. doi: 10.3389/fpsyg.2015.01379

immediate feedback (e.g., given by a coach). They indicated that these activities are not typically enjoyable and they distinguished deliberate practice from other activities such as work and play. The deliberate practice framework includes the strong statement that genetic differences among individuals do not explain differences in expert performance (except for the case of height in some sports such as basketball), and that genetic differences may only contribute to expert performance indirectly through deliberate practice (i.e., there may be genetic differences in the willingness to engage in long periods of deliberate practice).

As indicated by Campitelli and Gobet (2011) and Hambrick et al. (2014a) the deliberate practice framework claims that abundant deliberate practice is both a necessary and a sufficient condition to achieve high levels of expert performance in sports, games, arts, and science.

Answering Research Questions with Measures of Variability

In a study conducted with 104 chess players (see Gobet and Campitelli, 2007; Campitelli and Gobet, 2008 for details), among other questions, Campitelli and Gobet requested participants to indicate the number of hours of individual and group practice they had engaged in since they started playing chess. The procedure was similar to the one used by previous researchers who mostly favor the deliberate practice framework (e.g., Charness et al., 1996, 2005).

In order to test the research question "Is abundant deliberate practice a *sufficient* condition to achieve high levels of expert performance in chess?" Campitelli and Gobet (2011) reviewed previous literature on chess expertise and utilized three procedures. In this section I focus on one of them: calculating the variability of the number of hours to achieve the master level—a level of expertise 3.5 standard deviations higher than the mean[1]. If the variability is small, this would give support to the deliberate practice framework whereas a large variability would provide evidence against that framework. This procedure was based on Gobet and Campitelli's (2007) dataset. Gobet and Campitelli had access to archival data that allowed them to determine the exact year in which the players achieved the master level. They used these data in combination with the number of hours of practice that each player accumulated until they achieved the master level. They then calculated the variability on the number of hours required to achieve that level. They found a range from 730 to 16,000 h of individual practice to achieve the master level. Thus, the deliberate practice framework's hypothesis that abundant deliberate practice is a sufficient condition to achieve high levels of expertise was not supported by the data.

[1]The chess international rating system uses the Elo (1978) scoring system, which follows a normal distribution with a theoretical mean of 1500 and standard deviation of 200, in which the current world champion possesses a score of 2876. The psychology of chess literature typically uses the following hierarchy to categorize chess players: >2600 = grandmaster, >2400 = international master, >2200 = Master, >2000 = Expert or candidate master, >1800 = Class A player, >1600 = Class B player, >1400 = Class C player.

Answering Research Questions with One Value

As indicated by Campitelli and Gobet (2011) another way of testing the above hypothesis is to find one individual who engaged in abundant deliberate practice and failed to attain the master level. This would rule out abundant deliberate practice as a sufficient condition to achieve high levels of expert performance. Campitelli and Gobet reported that there were several players dedicating more than 20,000 h to chess who did not achieve the master level; therefore, the hypothesis that deliberate practice is a sufficient condition was not supported by the data.

Answering Research Questions with One Case

Ericsson et al. (1993) hypothesized that 10 years of intense dedication to a field are *necessary* to achieve high levels of expert performance. This claim was slightly changed and popularized to the general public by the writer Malcom Gladwell in his bestseller "Outliers" (Gladwell, 2008). Appealing meritocratic values Gladwell captured the public imagination by coining the "10,000 h rule": 10,000 h of intense dedication are *necessary* to achieve high levels of expert performance.

In order to test this hypothesis is not even necessary to collect data because archival data are available. Finding one case in which a high level of expert performance in chess is achieved in less than 10 years—in other words, finding a Mozart of chess–would refute the hypothesis. Indeed, Gobet and Campitelli (2007) identified more than one case: Ukranian Ruslan Ponomariov and Hungarian Peter Leko attained the grandmaster level (i.e., 2 levels [or 2 standard deviations] up the master level) at the age of 14, and in interviews they both reported having started playing chess at the age of 7. More impressively, Ukranian-born Russian Sergei Karjakin obtained the grandmaster level at the age of 12 and the international master level at the age of 11. At the age of 11 he was hired by Ponomariov to assist him in the preparation for the 2002 Chess World Championship match. More recently, the current world champion, Norwegian Magnus Carlsen obtained the grandmaster level at the age of 13 and reported that he played his first chess tournament at the age of 8 (see Gobet and Ereku, 2014, for more details on the case of Magnus Carlsen). Nowadays, there are 23 players who achieved the grandmaster level before the age of 15. These data suggest that 10 years or 10,000 h of intense dedication are not *necessary* to achieve high levels of expert performance.

Answering Research Questions with Distributions

Hambrick et al. (2014a) re-analyzed Gobet and Campitelli's (2007) data and presented a figure (see Figure 2, p. 39) with a distribution of hours of deliberate practice in three groups of chess players: master players, expert players and intermediate players (i.e., players with less than 2000 points ranging a number

of categories). Although the mean hours of deliberate practice between groups differ [master $M = 10{,}530$ h ($SD = 7414$), expert $M = 5673$ h ($SD = 4654$), and intermediate $M = 3179$ h ($SD = 4615$)], as suggested by the large standard deviations, the overlap among the three distributions is evident by just visual inspection. For example, as expected, more than 60% of the intermediate players practiced between 0 and 2500 h. If abundant practice were a necessary condition to achieve high levels of expertise it is not expected to have players of the other groups in this interval of low practice. However, more than 25% of the expert players and more than 10% of the master players are in this interval. Moreover, the mode of the master and the expert groups is located in the same interval (i.e., between 5000 and 7500 h of practice), with more than 30% of experts, almost 25% of masters and almost 10% of intermediate players located in this interval. Furthermore, as expected, about 25% of the masters accumulated more than 17,500 h of deliberate practice; but, unexpectedly, about 2% of the experts and about 3% of the intermediates also accumulated more than 17,500 h of deliberate practice.

Conclusion

As indicated by Speelman and McGann (2013), calculating a mean of some dependent variable as a first step of other statistical procedures is only one of a range of procedures available for the psychological researcher. There are two main reasons why psychological researchers tend to overlook the type of analyses presented above. First, psychological researchers are trained in application of statistical procedures that are typically useful for most types of research. Based on my experience with colleagues of other disciplines, this training is of high quality, thus psychological researchers have reasons to be proud of their analytic skills. However, the training focuses on the application, not the understanding, of those procedures. Indeed, research has shown that psychological researchers have difficulties in understanding p values (e.g., Badenes-Ribera et al., 2015). Second, psychology has a shortage of formal (i.e., mathematical, computational) theories that allow researchers to make precise numerical predictions of values (or a range of values) in experiments. This leads to relying on qualitative predictions (i.e., a group will have a higher average than another group) in which procedures involving calculating group means are the most appropriate. In this respect, Ericsson and colleagues should be credited for providing numerical predictions (i.e., 10 years (or 10,000 h) of deliberate practice are necessary to achieve high levels of expert performance), which can be tested with the analytic procedures explained above.

This article builds upon Speelman and McGann's (2013) call for critical use of statistical procedures, and illustrates four sound procedures to answer research questions, which do not involve calculating the mean. It is to be hoped that this article contributes toward the development of formal theories and ingenious procedures to answer research questions, as opposed to fitting research questions to the requirements of extant popular statistical procedures.

References

Badenes-Ribera, L., Frías-Navarro, D., Monterde-i-Bort, H., and Pascual-Soler, M. (2015). Interpretation of the p value: a national survey study in academic psychologists from Spain. *Psicothema* 27, 290–295. doi: 10.7334/psicothema2014.283

Campitelli, G., and Gobet, F. (2008). The role of practice in chess: a longitudinal study. *Learn. Individ. Differ.* 18, 446–458. doi: 10.1016/j.lindif.2007.11.006

Campitelli, G., and Gobet, F. (2011). Deliberate practice: necessary but not sufficient. *Curr. Dir. Psychol. Sci.* 20, 280–285. doi: 10.1177/0963721411421922

Charness, N., Krampe, R. T., and Mayr, U. (1996). "The role of practice and coaching in entrepreneurial skill domains: an international comparison of life-span chess skill acquisition," in *The Road to Excellence: The Acquisition of Expert Performance in the Arts and Sciences, Sports and Games,* ed K. A. Ericsson (Mahwah, NJ: Erlbaum), 51–80.

Charness, N., Tuffiash, M., Krampe, R., Reingold, E., and Vasyukova, E. (2005). The role of deliberate practice in chess expertise. *Appl. Cogn. Psychol.* 19, 151–165. doi: 10.1002/acp.1106

Cohen, J. (1994). The earth is round ($p < .05$). *Am. Psychol.* 49, 997–1003. doi: 10.1037/0003-066X.49.12.997

Elo, A. E. (1978). *The Rating of Chess Players. Past and Present*. New York, NY: Arco.

Ericsson, K. A. (2014). Why expert performance is special and cannot be extrapolated from studies of performance in the general population: a response to criticisms. *Intelligence* 45, 81–103. doi: 10.1016/j.intell.2013.12.001

Ericsson, K. A., Krampe, R. T., and Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychol. Rev.* 100, 363–406. doi: 10.1037/0033-295X.100.3.363

Gladwell, M. (2008). *Outliers: The story of Success*. New York, NY: Little, Brown and Co.

Gobet, F., and Campitelli, G. (2007). The role of domain-specific practice, handedness, and starting age in chess. *Dev. Psychol.* 43, 159–172. doi: 10.1037/0012-1649.43.1.159

Gobet, F., and Ereku, M. H. (2014). Checkmate to deliberate practice: the case of Magnus Carlsen. *Front. Psychol.* 5:878. doi: 10.3389/fpsyg.2014.00878

Hambrick, D. Z., Altmann, E. M., Oswald, F. L., Meinz, E. J., Gobet, F., and Campitelli, G. (2014b). Accounting for expert performance: the devil is in the details. *Intelligence* 45, 112–114. doi: 10.1016/j.intell.2014.01.007

Hambrick, D. Z., Oswald, F. L., Altmann, E. M., Meinz, E. J., Gobet, F., and Campitelli, G. (2014a). Deliberate practice: is that all it takes to become an expert? *Intelligence* 45, 34–45. doi: 10.1016/j.intell.2013.04.001

Speelman, C. P., and McGann, M. (2013). How mean is the mean? *Front. Psychol.* 4:451. doi: 10.3389/fpsyg.2013.00451

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Campitelli. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

METHODS
published: 24 July 2015
doi: 10.3389/fpsyg.2015.01007

From means and variances to persons and patterns

*James W. Grice**

Department of Psychology, Oklahoma State University, Stillwater, OK, USA

A novel approach for conceptualizing and analyzing data from psychological studies is presented and discussed. This approach is centered on model building in an effort to explicate the structures and processes believed to generate a set of observations. These models therefore go beyond the variable-based, path models in use today which are limiting with regard to the types of inferences psychologists can draw from their research. In terms of analysis, the newer approach replaces traditional aggregate statistics such as means, variances, and covariances with methods of pattern detection and analysis. While these methods are person-centered and do not require parametric assumptions, they are both demanding and rigorous. They also provide psychologists with the information needed to draw the primary inference they often wish to make from their research; namely, the *inference to best explanation*.

Keywords: observation oriented modeling, integrated model, inference to best explanation, mean, variable-based modeling

OPEN ACCESS

Edited by:
Craig Speelman,
Edith Cowan University, Australia

Reviewed by:
David Trafimow,
New Mexico State University, USA
Brian D. Haig,
University of Canterbury, New Zealand

***Correspondence:**
James W. Grice,
Department of Psychology, Oklahoma State University, 116 North Murray, Stillwater, OK 74078, USA
james.grice@okstate.edu

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *02 June 2015*
Accepted: *03 July 2015*
Published: *24 July 2015*

Citation:
Grice JW (2015) From means and variances to persons and patterns. Front. Psychol. 6:1007. doi: 10.3389/fpsyg.2015.01007

Introduction

In his erudite and now classic book *Constructing the Subject*, Danziger (1990) describes how psychology came to be dominated by an approach toward data conceptualizing and analysis he dubbed the "triumph of the aggregate." Charting the meteoric rise of tables of means, variances, correlations, and other aggregate statistics reported in psychology journals from the early to mid-1900s, Danziger (1990) lamented the corresponding demise of the individual subject, or person, in psychology. Aggregate statistics were moreover shown to rise in prominence despite cautionary claims regarding their hegemony (Sidman, 1952; Bakan, 1954), including a critical appraisal offered by none other than Skinner (1956) himself. Modern scholars point out the issues raised over 50 years ago have not gone away and that, in fact, psychology's over-reliance on aggregate statistics is likely thwarting scientific progress by hindering the development of theories which can explain the behavior of individual persons (Valsiner, 1986; Valsiner et al., 2014). Lamiell (1997, 2003, 2013) has gone to great lengths to remind personality psychologists, in particular, that between-person differences or effects discovered through aggregate statistical analysis do not necessarily exist at the level of the individual (see also, Carlson, 1971). The Big Five personality factors, for example, can readily be found in aggregated data, but the factors do not regularly emerge from the analysis of individual responses (Grice et al., 2006; see also, Molenaar and Campbell, 2009). The Power Law of Learning is another example phenomenon that can be seen in the aggregate but not at the level of the individual, thus raising the question of whether or not it is truly a law (see Heathcote et al., 2000, as reported by Speelman and McGann, 2013). There is a genuine and potentially hazardous disconnect, then, between conclusions drawn from between-person, aggregate statistics and statements or theories meant to offer insight into the psychology of individual persons.

In this paper we present a framework for conceptualizing and analyzing data that does not rely on traditional aggregate statistics such as the mean, median, variance, covariation, etc. Instead, this approach—like Exploratory Data Analysis (EDA; Tukey, 1977; Behrens and Yu, 2003)—relies primarily upon techniques of visual examination to detect and explain dominant patterns within a set of observations. Going beyond EDA, however, this approach can incorporate patterns that are generated *a priori* on the basis of theory, thus promoting model building and development. It also synchronizes visual examination of the data with transparent analyses that (1) identify those individuals whose observations are consistent with the predicted or identified pattern, and (2) provide an index of a given pattern's robustness within a sample. This approach is generally referred to as *observation oriented modeling* (OOM; Grice, 2011, 2014), and we will compare and contrast its guiding principles and techniques with those of traditional statistics using a study and accompanying data that are contrived but nonetheless based on genuine psychological research. We will also draw, in part, upon Haig's (2005; 2014) *abductive theory of method* (ATOM) to argue that this approach provides the types of inferences psychologists normally seek from their data but are unable to make on the basis of traditional statistical analyses. The overall goal is to show that by departing from the modal research practice of modern psychology (Lebel and Peters, 2011), a novel and more rigorous path that does not confuse aggregates for persons may be paved for future researchers.

An Example Study in Rejection

Pick up almost any research paper on human psychology, and there you will find written in the Introduction statements about persons. You will not likely find statements about means, variances, or even covariances; although you might find descriptions of relationships between different attributes or qualities. Even these relationships, however, will be discussed in the context of living persons rather than aggregate statistics. Writing about rejection and interpersonal coping, for example, Ayduk et al. (2003) claim "...research suggests that people who fear and expect rejection employ to a greater degree both overt (i.e., verbal aggression) and covert (i.e., withdrawal, avoidance) negative coping strategies that ultimately undermine their significant relationships and their mental health" (p. 435). Here rejection and interpersonal coping are foremost recognized as universal features of human experience. It is indeed difficult to imagine any adult who could not recall an instance of being rejected by another person or recount a situation in life that was coped with in a negative, unfruitful manner. The authors moreover infer from previous research—naturally based on a limited number of individuals—that rejection and coping are causally connected. The inference is therefore from samples of persons to persons in general, as is clear with the authors' use of the word "people." Concomitant with this inference is the conclusion that rejection and coping are causally connected, not at the aggregate or even group level, but at the level of the person. For any given individual, then, the chronic expectation of rejection (likely developed from a history of being rejected by others) plays a causal role in the generation of negative coping strategies. How did the authors draw such inferences, and are these types of inferences truly warranted when made on the basis of results obtained from traditional, aggregate statistics? The answer to the second question is "no," and to understand how psychologists typically draw such conclusions from their analyses, we must work patiently and carefully through an example study.

Continuing with the topic at hand, rejection can be produced and studied in the laboratory by psychologists (e.g., Downey and Feldman, 1996; Ayduk et al., 2003). Consider, for instance, a male college student (viz., "the participant") who walks into a laboratory and is informed that he will be interacting, via the Internet, with another male student seated at another computer across campus. The participant is asked to provide a short biography to share and is then given a corresponding biography from his counterpart on the other side of campus. The biography presents a person who is kind and inquisitive and likely a pleasure to interact with in an informal social setting. After the participant reads the biography and prepares for the online interaction, the experimenter receives a phone call and informs the participant that his counterpart has now chosen not to participate in the online discussion and is instead withdrawing from the study. What is the participant to make of this unexpected decision? The experimenter's hope is that the student will in fact interpret the counterpart's decision as a rejection of the participant based on his shared biography. The experimenter moreover expects the participant to subsequently experience negative emotions, make negative self-attributions, and to form a negative inclination toward his rejecting counterpart. With the phone call and rejection completed, the experimenter then asks the participant to judge the counterpart on qualities such as intelligence, popularity, and friendliness using a 6-point rating scale. These ratings essentially provide the student with an opportunity to express his displeasure with the rejecting partner. After making his ratings, the participant is finally debriefed and informed of the deception; viz., no other student was involved in the study, and the biography and ostensible rejection were therefore not genuine.

Now imagine over the course of a semester eighty individual students walking into the psychologist's laboratory and being guided through these same procedures. With each and every student the experimenter's expectations will be the same, because within her mind is a model. Perhaps it is a model that is only crudely elaborated, but it is a model nonetheless that is meant to explain the thoughts, feelings, and behaviors of each individual student (person) in the study. What might this model look like? The most rigorous way to express the model is via a picture like that shown in **Figure 1**. Such pictures are referred to as *iconic* or *integrated* models because they provide a visual snapshot of the structures and processes, or causes and effects, at work in the laboratory (i.e., at work in the participant, experimenter, and setting) during the study. As can be seen on the left side of **Figure 1**, the model depicts two conditions in the study. The top part of the model, demarcated by the bold line, represents what takes place in the study as described in detail above. The bottom part of the model will be described later.

The model also demarcates three important points of interaction between the experimenter and participant. The purpose of the first interaction, from the perspective of the

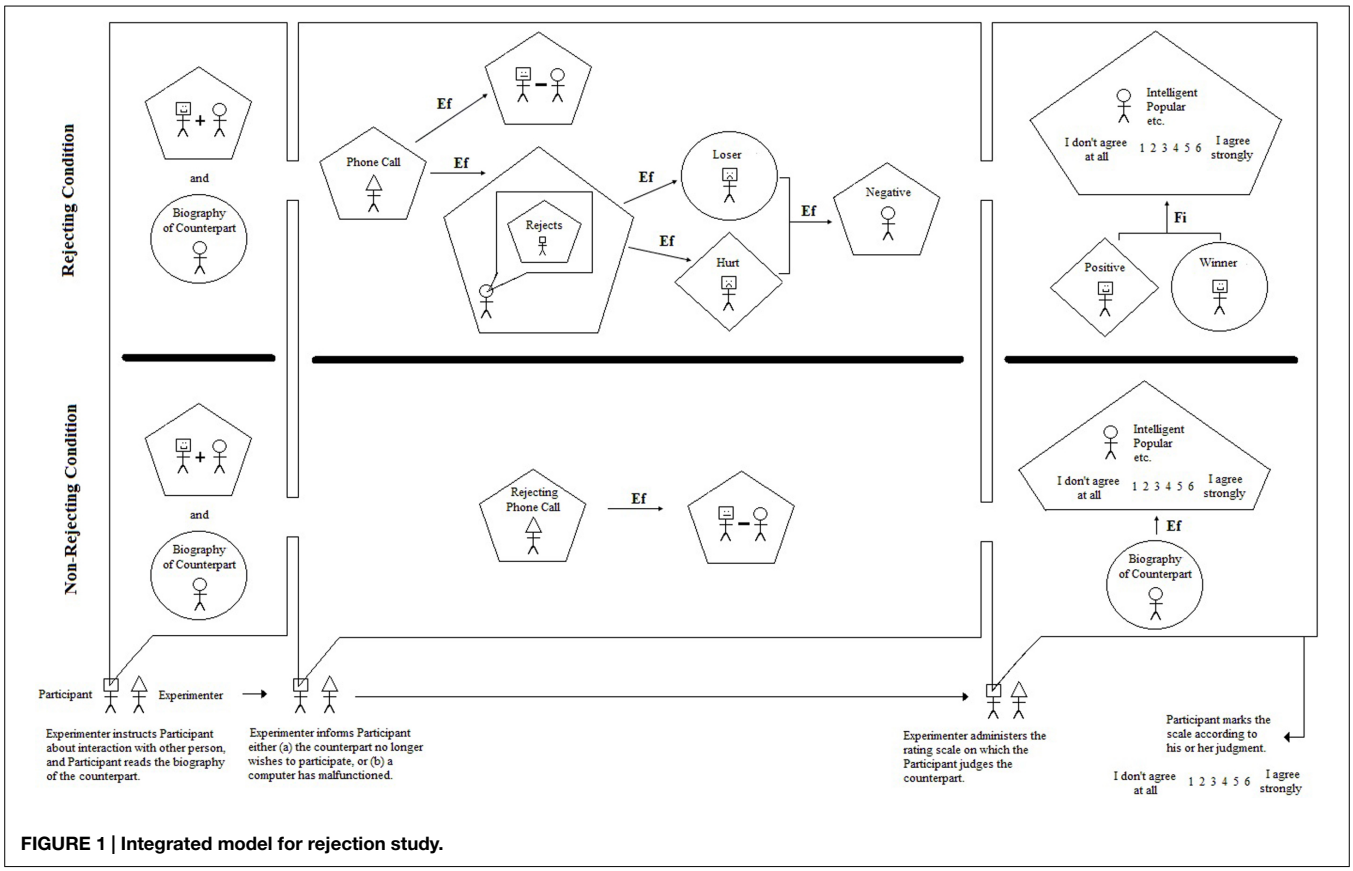


FIGURE 1 | Integrated model for rejection study.

experimenter, is to create within the mind of the participant an expectation of an online interaction with another student across campus. The pentagon enclosing the image of the participant and counterpart joined by a "+" sign represents a simple or complex judgment. In this instance, the participant judges that he will be interacting with the other student, and that the interaction will be positive (given the biography and the experimenter's instructions). The circle enclosing the image of the counterpart represents certain predicates (predicative adjectives, predicative nouns) based on the biography. For example, the biography describes the counterpart as "a student," "a psychology major," "outgoing," etc. The counterpart student is thus known through the neutral and positive descriptive nouns and adjectives given in the biography.

The purpose of the second interaction (focusing on rejecting the condition) is to inform the participant that the other student has chosen not to participate in the discussion after having read the participant's biographical statement. It is not perfectly clear or stated plainly, however, if the counterpart is rejecting the participant, but it is the experimenter's expectation that the participant will interpret this decision as a personal rejection based on his own biographical statement. The phone call is therefore considered to be an efficient cause; that is, a cause that proceeds its effect in time leading to its production or change (denoted by an arrow labeled "Ef" in the model; see Grice, 2014). The resulting judgment of the counterpart rejecting the participant is represented by the pentagon in the second interaction of the study. It is also accompanied by a second effect of the phone call; specifically, the simple judgment that the two will not communicate via the Internet after all.

The judged rejection then operates as an efficient cause of negative self-predications (negative self-attributions) by the participant. These negative self-predications are represented by the circle derogatorily labeled "Loser" enclosing the participant in the second interaction of the study. Hurt feelings, represented by a diamond labeled "Hurt" enclosing the participant, also result from the judged rejection. Finally, these experiences occurring simultaneously within the participant cause him (as an efficient cause) to adopt a negative disposition toward the counterpart. This negative disposition may occur consciously, for instance, if the participant were to think disparaging thoughts such as "well, that guy's a jerk for wasting my time" or "I always knew psychology majors were unstable."

Finally, the purpose of the third interaction is to provide the participant with an opportunity to make explicit judgments about his rejecting counterpart. As can be seen on the right side of **Figure 1**, the participant rates his counterpart using a 6-point scale anchored by "I don't agree at all" and "I agree strongly." The participant is asked to use the scale for nine adjectives that may describe the counterpart (intelligent, popular, friendly, etc.). Use of the scale is considered to be a complex judgment task as indicated by its enclosure in a pentagon. What value, or potential range of values, should the participant choose? At this point, the model is not sufficiently developed to specify exactly which values will be chosen, but it does explicate the most proximate cause behind the selection. Specifically, the participant

will choose a value that will result in positive feelings and positive self-attributions, as indicated by the diamond labeled "Positive" and the circle labeled "Winner" in the figure. He is therefore attempting to reach a goal through his rating, and goals operate as final causes in human behavior (see Rychlak, 1988; Grice, 2014). The arrow therefore points from the positive feelings and positive self-attributions and is labeled "Fi" for "final cause." Here the participant is essentially trying to make himself feel better through the rating judgment by discounting the source of his negative feelings and negative self-attributions (viz., the counterpart). It is therefore reasonable to posit that the participant will choose one of the low values on the scale (1, 2, 3), which would ostensibly indicate negative judgments of unintelligent, unpopular, unfriendly, etc.

At the end of the model (bottom right-hand corner of **Figure 1**) is the only "output" the experimenter observes; namely, the circled ratings for each adjective. No other attempts are made in the study to observe the predications, judgments, or feelings of the participant. Nonetheless, the figure spells out very clearly, for everyone to see, the structures and processes thought to be at work by the experimenter when the participant is ostensibly rejected. Finally, the model also shows a second condition of the study in which the participant is told that, due to a computer malfunction, the counterpart will not be able to take part in the discussion. The participant is still asked to rate his counterpart, but as can be seen in the model, all of the important predications, judgments, feelings, and causes are no longer present. His rating is driven by the biographical sketch, as an efficient cause, remembered from the beginning of the study. In this case, it would be reasonable to posit that the participant will choose one of the high values on the scale (4, 5, 6) due to the positive content of the sketch. Again, for both the rejecting and non-rejecting conditions, however, predictions for the scale values are not explicitly provided by the model.

Three Important Inferences

The integrated (iconic) model in **Figure 1** facilitates three inferences the psychologist wishes to make through her research efforts—even if she is not consciously aware of these inferences—and they are the types of inferences described at the beginning of this paper. The first is known as an *abductive inference* (or simply *abduction*) which has its roots in Aristotelian philosophy and was developed and popularized by the American philosopher Charles S. Peirce (Haig, 2005; Flórez, 2014). In order to understand this inference, let us suppose that all of the participants in the rejecting condition selected "I don't agree at all" for each and every adjective as applied to the counterpart. The data obtained from the study are recorded as whole numbers valued 1 through 6, and here all of the observed values for the rejected participants are 1's. Why do the numbers show this striking pattern? The experimenter's answer to this question, that is her conclusion or inference, is that the data are patterned in this manner because of the structures and processes detailed in the model. The specific form of this abductive inference can be represented as follows:

1's have been observed for the rejected participants.

If the structured processes in **Figure 1** took place, 1's would have been observed.

Therefore, the structured processes in **Figure 1** took place.

A key feature of the inference is that, unlike induction, it appeals to explanation (viz., the structures and processes diagrammed in **Figure 1**). The conclusion is also uncertain or provisional, unlike conclusions reached through strict deductive reasoning. This uncertainty rests partly upon the iconic model itself as it is not sufficiently developed to predict that only 1's will be selected by the participants. Moreover, it is not clear if 1's should be selected for all nine adjectives, or if 2's and 3's are also expected since they lie below the mid-point of the scale, ostensibly conveying a negative judgment. Indeed, it is not clear why a 6-point rating scale is being used rather than, for instance, a 5-point scale or a binary judgment task. These uncertainties are part and parcel of the inference being sought and they should not be viewed as reasons for abandoning the explanatory model. Instead, these insufficiencies should be viewed as a call to make improvements to the model in **Figure 1** through refinement and extension.

The model can be refined by changing its existing components; for example, the exact emotions felt by the participant can be elaborated, or the 6-point scale can be justified and predictions included about how the participant should behave with regard to the scale. The model can be extended by adding additional components; for example, perhaps not every participant will construe the counterpart's actions as rejection, and the determining factors for making such an interpretation can be added to the model. Of course the entire model itself can be tested against and perhaps superseded by a competing model (such as a Freudian view of hostility). In this regard, in particular, the experimenter is seeking an *inference to best explanation*, which is a type of scientifically useful abduction with the general form,

D is a collection of data

H (an hypothesis) would, if true, explain D

No other hypothesis can explain D as well as H does

Therefore, H is probably true

The conclusion is again uncertain, but the continual striving to thoroughly evaluate, improve upon, or replace a given model seems to capture the investigative spirit of modern science, at least as it is idealized. In any case it is easy to see why Haig (2014), in his ATOM, regards inference to best explanation as central to developing a proper understanding of science.

The third inference sought by the experimenter is an attempt to draw a conclusion about persons, in general, from her specific sample of individuals in her particular study. The model in **Figure 1** is clearly tied to the study designed by the experimenter and is therefore applicable, she hopes, to each and every participant in her study. Beyond these persons, however, the components of the model provide a potential explanation for how persons, in general, might react to being rejected. The second interaction between the experimenter and participant, for instance, shows a causal link between the judgment of being

rejected by another person and the negative self-attributions and feelings experienced by the participant. The third interaction moreover shows that the proximate cause of a negative, explicit judgment toward a rejecting person is a final cause resulting in positive feelings and self-attributions. Presuming the observations do in fact support the model in her study, the experimenter can then argue that these components of the model may offer valid explanations of how persons react to rejection in situations outside the laboratory. In doing so, the psychologist will be reasoning inductively, moving from the specific to the general, and it is in this way that psychologists typically seek to generalize beyond their samples of participants and particular laboratories.

The three types of inference sought by the experimenter in this example are therefore (1) abduction, (2) inference to best explanation, and (3) inductive generalization. All three inferences are facilitated by the integrated, iconic model in **Figure 1**; indeed, it could be argued that such models are indispensable for making these inferences. In any case the inferences are clearly important, and as noted at the beginning of this paper, they are the types of inferences encountered in the Introduction and Discussion sections of journal articles published throughout psychology. It is also important to point out that none of these inferences is tied explicitly to a mean, median, mode, variance, or any other aggregate statistic that can be computed from a sample of data. The integrated model was designed without any statistical procedure in mind and without the restriction of only including features that can be understood quantitatively. All 10 of Aristotle's categories of being, and all four of his causes can be incorporated into an integrated model (see Grice et al., 2012; Grice, 2014). The theoretical horse, so-to-speak, is therefore in front of the data analytic cart, as it should be. In more sophisticated language, we are not letting our methods determine our metaphysics (Rychlak, 1988).

One Underwhelming Inference

When psychologists argue they are using statistics to generalize beyond their samples, it is important to realize most believe they are generalizing in the manner described above; namely, making an abductive inference to best explanation or making an inductive inference about persons in general. Unfortunately, in the overwhelming majority of cases, nothing could be further from the truth. By using traditional statistical methods that rely on null hypothesis significance testing (NHST; viz., traditional *p*-values), psychologists are instead routinely making an *inference to a population parameter*, which is far less informative and far less useful for building scientific theories than the three inferences drawn from integrated models described above.

To demonstrate this ubiquitous type of inference, let us now consider the condition in which the participant is told that his counterpart cannot participate in the online discussion because of a computer malfunction. With this comparison group in place, and following modal research practice (Lebel and Peters, 2011), the experimenter now thinks about the study using the variable-based model in **Figure 2**. As can be seen, this model is comprised of an independent variable (viz., group) and dependent variable (viz., popularity rating) connected with a line that represents their relationship or correlation. The negative sign above the

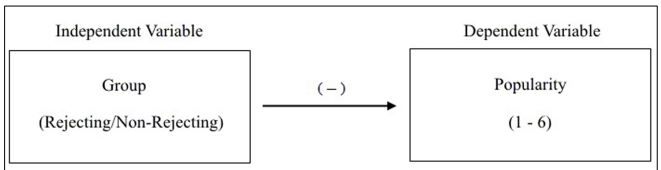


FIGURE 2 | Variable-based model for rejection study.

line indicates that those in the rejecting condition are expected to, on average, provide lower ratings than those in the non-rejecting condition. In order to keep everything simple, we will henceforth only consider the rating for "popular" in the analyses. Given the dichotomous group membership variable and rating scale with values ranging from 1 to 6, the experimenter follows standard protocol and analyzes the data with an independent samples *t*-test. Her results, obtained from 160 participants, reveal a statistically significant difference between the rejecting ($M = 4.20$, SD $= 0.40$) and non-rejecting ($M = 4.50$, SD $= 1.21$) groups, $t(96.23) = -2.10$, $p < 0.04$, $d = 0.33$, $CI_{0.95}$: -0.58, -0.02. The difference is also consistent with expectation, with the rejecting group yielding a lower mean than the non-rejecting group.

What inference can she draw from these results, assuming she has met or properly adjusted for all of the assumptions of the statistical test? Having used NHST, the experimenter posited two populations from which she drew her samples: a population of persons experiencing rejection in the study, and a population of persons not experiencing rejection in the study. The populations in this example, as in most studies in psychology, are entirely imaginary (Berk and Freedman, 2003); but nonetheless a mean rating value is presumed to exist for each, designated as μ_1 and μ_2. The null hypothesis is that the two population means are equal (H_0: $\mu_1 = \mu_2$) and by declaring her results as "statistically significant" she has rejected this hypothesis and concluded (inferred) that the two population means are not equal. She can consider the difference between the population means as a parameter to be estimated as well (viz., $\mu_{diff} = \mu_1 - \mu_2$) and then provide a point estimate for what she thinks the difference might be (viz., $4.20 - 4.50 = -0.30$). She can also provide an interval with an assigned level of confidence for possible values of the difference (viz., $CI_{0.95}$: -0.58, -0.02).

With the point and interval estimates in hand, it is clear the psychologist is attempting to make an inference to a population parameter (μ_{diff}), which is presumably fixed at some value. This inference is the only one she can rationally make; and the term "rationally" should be used loosely here because the low observed *p*-value ($p < 0.04$) does not provide the probability of the null hypothesis being true, and therefore worthy of rejection. The *p*-value instead indicates the two-tailed probability of obtaining a *t*-value of at least 2.10, assuming the hull hypothesis is true (see Cohen, 1994). Regardless, she cannot make an abductive inference due to the simplicity and nature of the variable-based model. The model is not explanatory as it does not present the structures and processes underlying the observations. It simply conveys the mean difference between two variables for arbitrarily defined populations. According to Haig's ATOM such a model

TABLE 1 | Statistics and independent samples *t*-test results for three studies.

	Condition								
	Rejecting		Non-rejecting						
Sample	*M*	SD	*M*	SD	M_{diff}	*d*	*t*	*p*	$CI_{0.95}$
1	4.20	0.40	4.50	1.21	−0.30	0.33	−2.10	0.037	−0.58, −0.02
2	4.20	2.24	4.80	1.34	−0.60	0.33	−2.06	0.042	−1.18, −0.023
3	4.20	0.89	4.50	0.93	−0.30	0.33	−2.10	0.038	−0.58, −0.02

The t-values and p-values for Samples 1 and 2 were adjusted for violations of the homogeneity of population variances assumption. All sample sizes were equal to 80.

with the accompanying parameter estimation may contribute to phenomena detection, which can play an important role in science, but the psychologist must be clear that the only conclusion she can draw from her analysis is that, provided the assumptions for the independent samples *t*-test have been met or adjusted for appropriately, the mean population difference is not 0, consistent with expectation, and is instead estimated with 95% confidence to be encompassed by values ranging from -0.02 to -0.58. That is all.

The experimenter also cannot make an inductive inference to people in general as her hypotheses and analysis are constrained to means. She cannot, therefore, write statements such as "people who are rejected will rate the rejecting person as less popular than those who are not rejected" or "rejected persons, compared to non-rejected persons, considered the counterpart to be unpopular." In order to be true to her model and analyses, she must restrict her inferential statements to population means or the difference between them. Moreover, she must be careful to avoid the following erroneous conclusions from her statistically significant finding:

- Because my result was statistically significant, it will likely replicate across independent samples of participants.
- My result is not likely due to chance given the low *p*-value.
- The null hypothesis is probably false; that is, the probability the null hypothesis is true is less than five percent.
- My research hypothesis is probably true.

Lambdin (2011) reports a more complete list of twelve such errors commonly made by researchers in psychology, education, sociology, medicine, and other disciplines who rely on null hypothesis significance testing (i.e., common *p*-values) to determine the scientific value of their results.

Getting Beyond Aggregate Statistics and NHST

A side-by-side comparison of the models in **Figures 1** and **2** shows clearly the integrated model is much more informative and rigorous than the variable-based model. The arguments above have also shown that the integrated model provides a gateway for the experimenter to make the types of inferences she truly wishes to make, whereas the variable-based model permits only a restricted, low information inference to a population parameter. In order to drive home the point that the latter inference is low in information, let us consider two additional

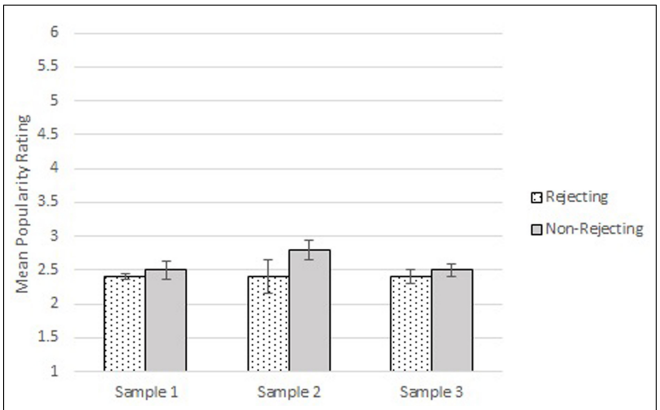


FIGURE 3 | Means and standard errors for rejecting and non-rejecting groups for each of the three samples.

samples of participants collected by the same experimenter using the exact same experimental protocol with a rejecting and non-rejecting condition. The descriptive statistics, *t*-values, *p*-values, and confidence intervals for all three samples are reported in **Table 1**, and the means and standard errors are displayed in bar graphs in **Figure 3**. As can be seen in the table, using these metrics she may conclude that the initial results have been replicated in the two new studies. The effect sizes, in particular, are equal ($d = 0.33$) when reported with two decimals of precision.

What do we really know about these data based on **Table 1** and **Figure 3?** Simple bubble plots surprisingly indicate that important information has been overlooked by focusing only on the tabled statistics. The bubble-plot in **Figure 4** for the first and original data set shows radical differences between the two groups with respect to the variability and distributions of their scores. While an overwhelming majority of participants in the rejecting group chose 4's, participants in the non-rejecting condition chose values ranging from 2 to 6. The second bubble-plot indicates a radical divide in the distribution for the rejecting group, with participants choosing only 1's, 2's, or 6's; whereas the distribution for the non-rejecting group shows skew toward the lower values on the scale. Finally, the third bubble-plot indicates the observed values are distributed similarly across the 6-point scale, with a slight tendency for participants in the rejecting group to select 3's and a slight tendency for participants in the non-rejecting group to select 6's.

The results from the three studies clearly show different patterns of responses that are simply not detectable in the

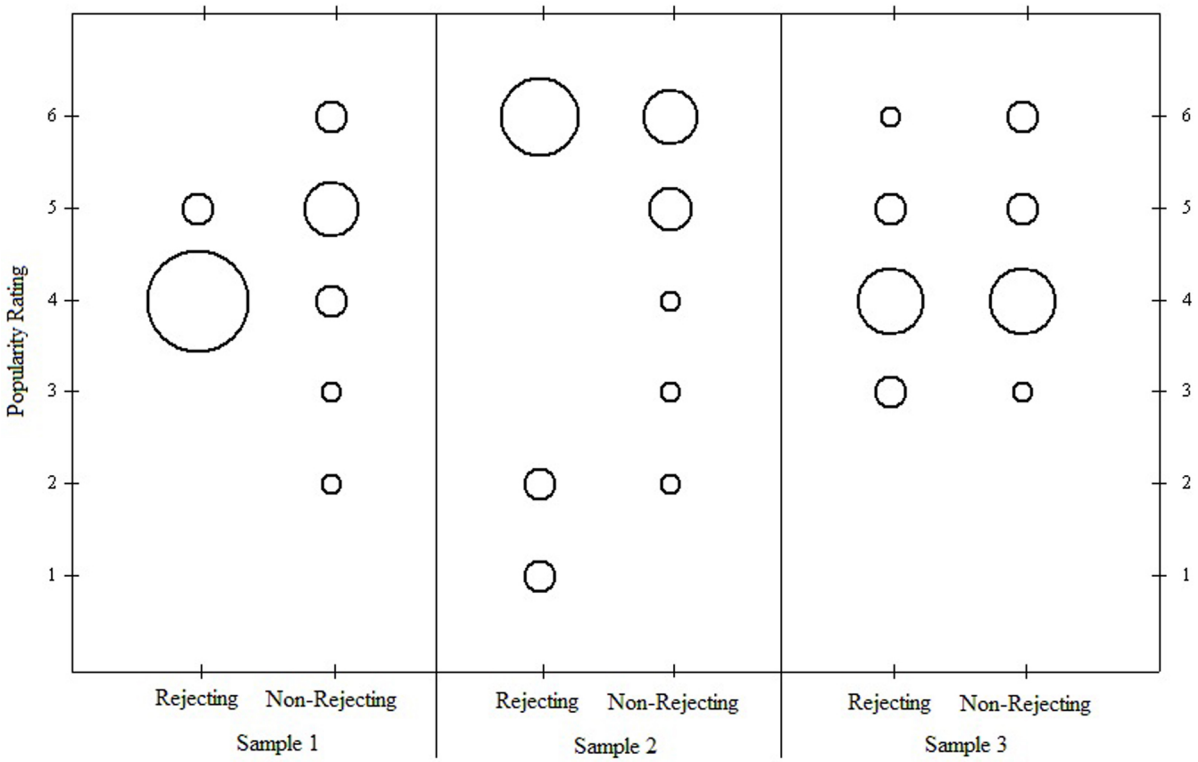
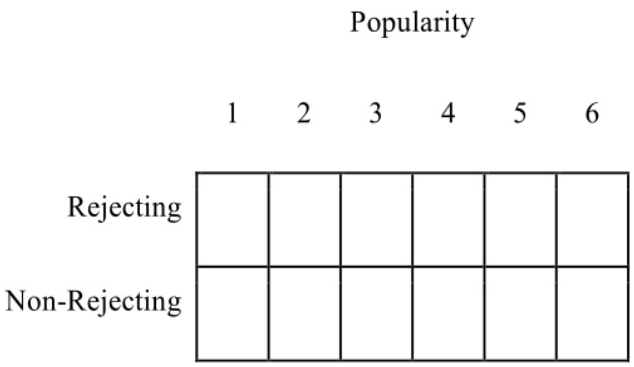


FIGURE 4 | Bubble plots for three samples of participants. Larger bubbles indicate greater numbers of cases.

aggregate statistics or bar charts. What is the experimenter to do? She could switch to a non-parametric procedure, but there are clear incentives for not doing so, including the potential loss of statistical power and the unwarranted perception that such a switch would indicate weakness in her methods and results. A median test in fact yields statistical significance for only the first two data sets. She could switch to a Bayesian analysis which would permit her to compare means while also assessing parameters relevant to the distributions of the samples. For all three data sets the Bayesian analysis in fact indicates "credible differences" between the group means, as the Highest Density Interval excluded 0 in each case. Fundamentally, though, none of these options represents a departure from the variable-based model in **Figure 2** and the goal of estimating parameters. In other words, like the independent samples *t*-test, effect size, and confidence interval these approaches would not require nor encourage the experimenter to explicate the structures and processes at work in or outside of her laboratory regarding the human experience of rejection.

The first step toward a more rigorous analysis of the data that is also consistent with the types of inferences sought through the model in **Figure 1** is to consider the detection and explanation of patterns as more generally important than parameter estimation (see Thorngate, 1986; Manicas, 2006). The experimenter has two key observations for each participant: (1) whether or not the participant was rejected, and (2) the participant's ratings using the 6-point scale. Here, as above, we will only consider the rating for popularity, and the two observations together create a simple two-dimensional array:



Given the experimenter's choices, then, this array presents boundaries on the ways she thinks data can be structured, and it is within this limiting structure she must identify or search for meaningful and robust patterns of observations.

Deductive, *a priori*, Pattern Evaluation

If the model in **Figure 1** were sufficiently developed, the experimenter would approach her data in a way most similar to deductive reasoning. In the parlance of modern research design and statistical analysis, she would conduct *a priori* tests of the model's accuracy which would require specific predictions about the observations. **Figure 5** shows two example predicted patterns

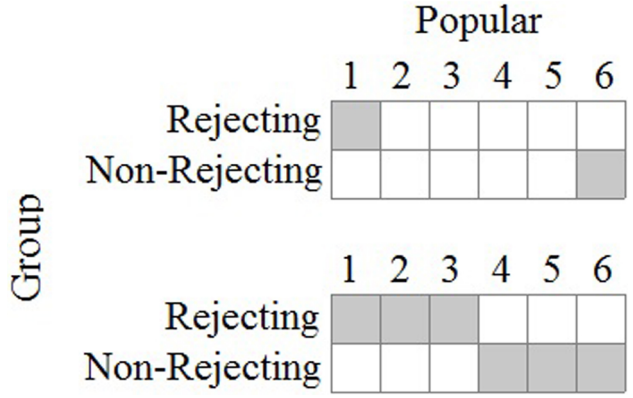


FIGURE 5 | Predicted patterns of observations for rejection study. Gray cells indicate predicted joint observations for group membership and popularity rating.

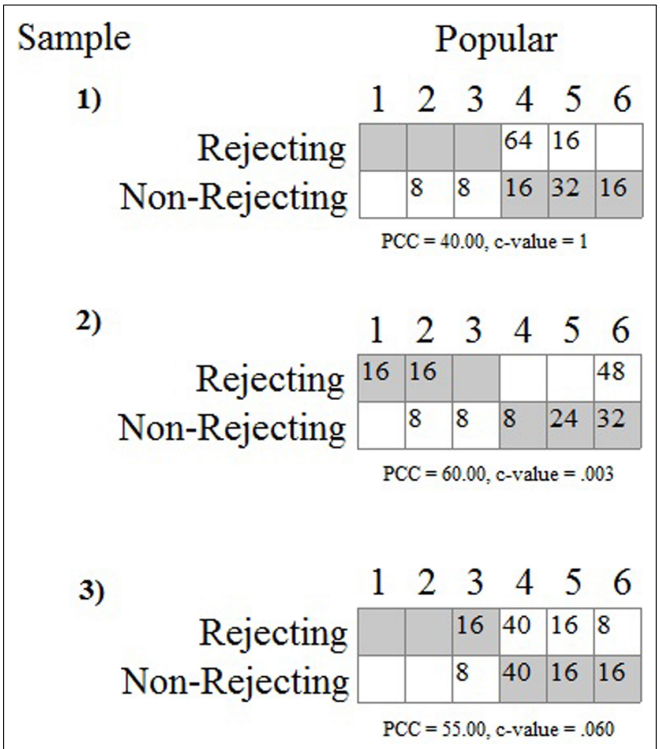


FIGURE 6 | Predicted pattern and actual observations for three samples from the rejection study. Gray cells indicate predicted joint observations for group membership and popularity rating.

using the two-dimensional array above that might be consistent with the integrated model. The first pattern shows that the experimenter expects participants in the rejecting group to select 1's ("I don't agree at all") on the 6-point scale and participants in the non-rejecting group to select 6's ("I agree strongly"). The second pattern shows that the rejected participants are expected to choose values 1, 2, or 3, while the non-rejected participants are expected to choose values 4, 5, or 6. These patterns are consistent with the model insofar as rejecting participants are expected to discount the counterpart, and lower values are interpreted as indicating negative judgments of low popularity.

Of course other patterns could be put forth as examples, but the point here is that if the experimenter is to work deductively and conduct *a priori* tests, she must develop the integrated model beyond what is shown in **Figure 1**. If she continues to employ a 6-point scale, she must be able to make predictions about which specific values will be selected by all—or at least a majority of—her participants. Such predictions will no doubt be difficult and will require extensive research into how individuals interpret and respond to the rating scale, but this is the demanding and often tedious scientific work required for accomplishing a better understanding of the scale values. By comparison, the variable-based model and independent samples *t*-test made few demands on the experimenter with regard to the meaning of the scale values, and they moreover required her to assume interval scale measurement and to assume that popularity itself is structured as a continuous quantity. No scientific evidence exists for either of these assumptions, and by thinking of her task as pattern identification the experimenter can avoid these assumptions while also pushing herself to think more deeply about what her numbers (i.e., the observations) actually mean.

For the sake of demonstration, let us assume that the second pattern in **Figure 5** is predicted by the integrated model. The actual observations from the three data sets can then be evaluated using the OOM software (Grice, 2011). The experimenter first sets up the two-dimensional array and defines the pattern. The frequencies are then computed and overlaid in the array, as shown in **Figure 6**. These are the primary results to be evaluated by the experimenter, and it can readily be seen that the observations from the first sample do not fit the pattern very well at all. None of the eighty participants in the rejected condition selected the 1, 2, or 3 values on the scale; and 16 participants in the non-rejecting condition selected these values, against expectation. Almost all of the 160 participants (90%) chose values of 4, 5, or 6. If these numerically high values are interpreted to represent the participant judging the counterpart as popular, and thus delivering a positive evaluation, then every person in the rejecting condition held a favorable attitude toward the counterpart.

Tallying all of the persons who were consistent with the predicted pattern yields what is known as the percent correct classification (PCC) index in the OOM software. The PCC index for the first sample was only 40%, as only 64 of the 160 joint group/rating observations matched expectation. The PCC index can range from 0 to 100 and is easily interpretable in light of **Figure 6**. A distribution-free randomization test can be conducted as an aid for interpreting the PCC index, the results of which are reported as a probability statistic known as the *c*-value (or *chance*-value). Relatively low values indicate the magnitude of the observed PCC index was not easily equaled or exceeded when computed from randomized pairings (1000 trials) of the group and popularity ratings for the 160 participants. For the first sample the *c*-value was 1.0 (possible range: 0–1), thus indicating that in every instance, the PCC index from randomized versions of the same data equaled or exceeded 40%. The observed PCC index was therefore not only low, but values at least that high were entirely ordinary as well.

The results in **Figure 6** for the first sample are in direct contradiction to any reasonable expectation based on the integrated model. Yet, recall from **Table 1** above the outcome from the *t*-test was statistically significant and interpreted as offering support for the variable-based model because the average rating for the rejecting group was lower than the average for the non-rejecting group. The place on the scale this difference occurred did not matter: the difference between 1 and 2 has the same meaning as a difference between 3 and 4 in the *t*-test analysis. With the OOM analyses, by contrast, the scale values had to be taken seriously when defining the expected pattern and interpreting the results.

The second data set also reveals striking results that were masked by the traditional statistics; specifically, as can be seen in **Figure 6**, 48 individuals (60%) in the rejecting condition rated the counterpart as a six on the 6-point scale. Again, these observations make no sense in light of the integrated model. The data for the rejecting condition are moreover split between the ends of the scale as the remaining 32 individuals selected 1's or 2's. Given this odd pattern the PCC index, which equals 60% and was unusual compared to randomized versions of the observations (*c*-value = 0.003, 1000 trials), is to be interpreted cautiously or even ignored. More specific analyses must also be conducted in this case, treating the two groups separately. The PCC index for the rejecting participants, treated separately, was only 40% (*c*-value = 0.98), while the PCC index for the non-rejected participants was impressively high (80%, *c*-value < 0.001). Expectations were therefore largely met for the non-rejected participants but not for the rejected participants.

Figure 6 shows the results for the third data set to be entirely unimpressive, even though the *t*-test was again statistically significant. As can be seen, a majority of the participants in the rejecting condition again chose 4, 5, or 6 from the rating scale. Equal numbers of participants in the non-rejecting condition chose 4's and 5's, and not a single individual from either group chose 1 or 2. The differences between the groups occurred only for values of 3 and 6, with more participants in the rejecting condition choosing 3's and more participants in the non-rejecting condition choosing 6's. The PCC index (55%) indicated that barely over half of the students were classified correctly which was even less impressive than the value obtained for the second data set, even though it was also unusual based on the randomization test (*c*-value = 0.06). Again, the results shown in **Figure 6** are primary, and as a general rule in OOM PCC indices and *c*-values should never be presented without clear visual displays of the data. Opposite of NHST, as well, the probability statistic (viz., the *c*-value) is the least important bit of information in the analysis, and in this particular set of analyses it may even be considered superfluous.

Abductive, *post hoc*, Pattern Evaluation

The model in **Figure 1** does not explicitly predict which values on the 6-point scale will be chosen by individuals in the two groups. Without such specificity in the model, the experimenter must approach the three data sets in a manner consistent with inductive and abductive reasoning. In the parlance of modern research design and statistical analysis, she must examine the data *post hoc* for robust and meaningful patterns. She can do so using the OOM software and what is known as binary Procrustes rotation, which is a procedure that seeks to rotate one set of observations into conformity with a second set of observations (Grice, 2011).

Results for the three data sets, displayed as multigrams, are shown in **Figure 7**. As can be seen for the first sample, the multigram is comprised of two aligned histograms for the rejecting and non-rejecting groups. The bars in the multigram are shaded or filled on the basis of the Procrustes rotation. A shaded bar indicates those observations that are considered correctly classified by the algorithm, while a bar filled with diagonal lines indicates those observations incorrectly classified. It is important to keep in mind that the analysis is entirely *post hoc*. The observations are classified as correct or incorrect by the rotation algorithm on the basis of the patterns of frequencies considered both between groups and across the six scale values. The experimenter in no way determines how the observations are expected to be patterned or considered as accurately or inaccurately classified. She must instead examine the pattern in the multigram and attempt to draw an inductive generalization and an abductive explanation.

The multigram for the first sample shows a convincing pattern with regard to the PCC index (80%). As can be seen in **Figure 7**, the largest bars in the multigram are shaded to indicate the correctly classified observations. The *c*-value from the randomization test is also impressively low. Not one time in 1000 trials did randomized versions of the actual observations yield a PCC index of 80% or more. The pattern is thus unusual, leading the experimenter to inductively reason that some phenomenon has potentially been detected. At the same time, the experimenter is confronted with the pattern in **Figure 7** and must work abductively to explain it. As can be seen, rejected participants were classified correctly only if they chose 4 on the 6-point scale. Non-rejected participants were classified correctly if they chose 2, 3, 5, and 6. How does the model in **Figure 1** comport with such a pattern? It is difficult to reconcile this observed pattern with the structures and processes in the integrated model, but the experimenter must try to do so. Alternatively, she can seek to modify the integrated model to explain the pattern. In either case, she is reasoning abductively as she ultimately seeks to make an inference to best explanation of the phenomenon she's detected through her study.

If the experimenter has all three data sets to work with, however, it is clear that a single pattern has not emerged. The multigram for the second sample in **Figure 7** shows that, contrary to the first sample, no participants in the rejecting group selected 4 from the scale; they selected only 1's, 2's, or 6's, with a majority choosing 6's. A majority of participants in the rejected group of the third sample chose 4's, but an equal number of participants chose 3's, 5's, or 6's. The distributions of observations across the six scale points for the non-rejected participants were, by comparison, more similar across all three samples, although the modes were different for each. The PCC index for the second sample was high (75%) and unusual (*c*-value < 0.001), whereas the PCC index for the third sample was low (20%) and easily equaled or exceeded by randomized versions of the data (*c*-value = 0.98). The pattern for the third sample also shows ambiguous classifications, indicating

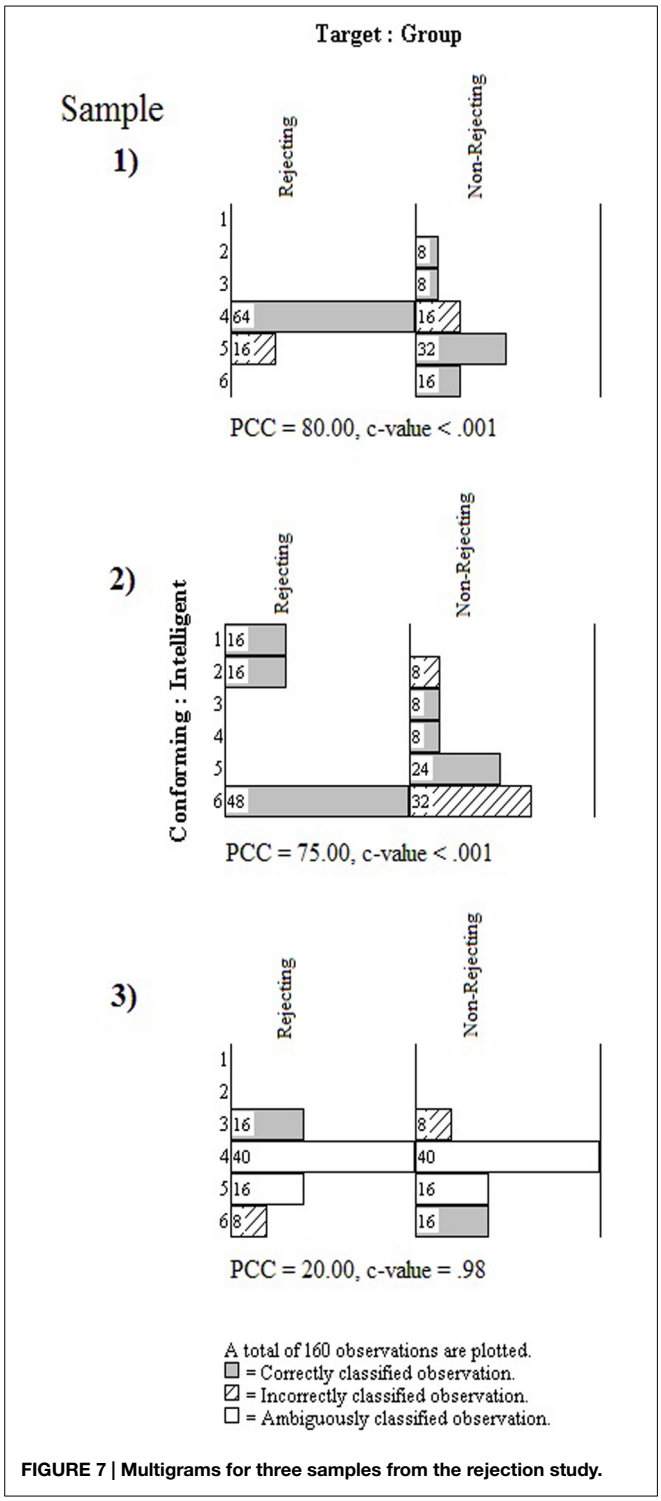


FIGURE 7 | Multigrams for three samples from the rejection study.

that the algorithm could not clearly distinguish between values of 4 and 5 for the rejected and non-rejected participants. Given the low PCC index and high *c*-value for this sample, the experimenter would interpret these results as not supporting the integrated model, and the data offer no clear pattern from which to generalize or alter the model. The relatively impressive individual results for the second sample would warrant further abductive attention. Considering all three sets of results together with their remarkably different patterns, however, may instead decide to conclude that no phenomenon has been reliably detected.

Discussion

Speelman and McGann (2013) report a number of assumptions about the mean held by modern psychologists. In light of the history, models, methodology and data analytic techniques examined in this paper, perhaps the most troubling assumption is that "any inability to use the mean as a reliable measure of a stable characteristic is a product of weakness in methodology or calculation" (p. 3). This assumption is disturbing for two reasons. First, how is it possible that one, simple statistic can be given so much power in the vast domain of scientific inquiry? Surely the spectacular advances in the fields of biology, chemistry, physics, and medicine, with all of their methodological rigor, have not depended on the lowly mean. The curious elevation of the mean in psychology as an indicator of rigor or as some type of "error free value"—or worse, "ideal person"—is the epitome of what Lamiell (2013) termed "statisticism." Philosophically, it is the error of placing methods before metaphysics; in other words, allowing methods of data collection and analysis to determine how one builds a model of nature. The practical result of such a limiting attitude is a guaranteed restriction in the advancement of psychological science.

Second, the idolatry of the mean is disturbing because it reveals that psychologists are operating under a quantitative imperative (Michell, 1999, 2008). What is popularity? What is rejection? What is the emotion of anger? Under the quantitative imperative the answer to each of these questions must in some way invoke the notion of continuous quantity. In other words, each of these qualities of human experience are presumed to exist in such a way as to be measurable as continuous quantities. In the parlance of Steven's (1946) four scales of measurement, popularity, rejection, and emotion must be measured as interval or ratio scales, for it is only with these types of scales that the computation of a mean is appropriate. Unfortunately, there is no evidence to date that qualities such as intelligence, depression, and personality traits (let alone popularity, rejection, or anger) are structured as continuous quantities, and therefore measurable as such. As stated by Michell (2011), "There is no evidence that the attributes that psychometricians aspire to measure (such as abilities, attitudes and personality traits) are quantitative" (p. 245). This is again an instance of putting methods ahead of metaphysics; that is, of presuming psychological qualities to be measurable as continuities without substantiating this claim and seriously considering the possibility that such qualities may be structured differently.

One need only examine the periodic table of the elements or the biochemical pathways of a eukaryotic cell to understand that the scientific study of nature is not restricted to interval and ratio scaled measurement and parametric statistics. The arguments, models, and methods, presented in this paper hopefully elucidate why psychologists should feel confident in venturing beyond the world of means, variances, and covariances without fearing a loss of scientific rigor. Placing the integrated model in

Figure 1 side-by-side with the variable-based model in **Figure 2** should be sufficient to convince the reader that theoretical rigor is in no way tied to an aggregate statistic of any kind. Many of the components in **Figure 1** (e.g., all of the acts of predication and most acts of judgment) are not even quantitative in nature, thus precluding the computation of a mean and variance. An integrated model like the one in this paper clearly requires a great deal more thought and effort to construct, validate, and defend than a variable-based model (see also Grice, 2011, 2014; Grice et al., 2012). Indeed, the reader is invited to sketch an integrated model for his or her most recent study, posited psychological process, or favorite theory. The task will no doubt prove challenging, but it will finally heed Meehl's (1978) call for more serious theorizing and bolder predictions in psychology. Not by coincidence, in the same paper Meehl argued that the over-reliance on null hypothesis significance testing was preventing scientific progress in psychology,

> "I believe that the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories in the soft areas is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology" (p. 817).

The shift to iconic modeling is also a step toward the types of inferences psychologists truly wish to make from their research: abductive inference, inference to best explanation, and inductive generalization. Variable-based models are meant to show associations between variables and are poor tools for explaining the complex structures and processes of nature. The mean does not provide information about "people in general" and in fact likely describes no one in particular (Lamiell, 2013). Variable-based models and their accompanying aggregate-based analyses are therefore not up to the task of delivering these inferences. When psychologists employ such methods and tie them to null hypothesis significance testing (traditional *p*-values), they are limited to drawing inferences about population parameters. . .regardless of whether or not they are cognizant of this fact. Using Haig's (2005) ATOM, these inferences may be of value insofar as they are seen as equivalent to phenomenon detection. The Flynn Effect, for instance, is the phenomenon of increased scores on intelligence tests over the past 30 years or so "detected" using aggregate statistics (Haig, 2014). The explanation of this phenomenon, however, will require a great deal more work and the construction of an integrated model that details the structures and processes underlying the Flynn Effect.

Going beyond the world of variable-based modeling and the computation of means, variances, and other parametric statistics is not necessarily a leap into the world of Bayesian statistics or non-parametric analyses; rather, the move is from estimating parameters in the context of sampling variability (as with an independent samples *t*-test) to the analysis of patterns in the context of explanatory models. Thorngate (1986) wrote plainly, "The essence of science is the detection and explanation of patterns" (p. 71), and he wrote this statement in a chapter for a book titled *The Individual Subject and Scientific Psychology* (Valsiner, 1986). Countless students have entered psychology expecting to study the lives of individuals only to learn that their task is instead to study variables, aggregates and some non-existent "average person." When collecting and analyzing data they learn that the odd person is a statistical nuisance or outlier who must be sacrificed to the mean or some statistical assumption (e.g., homogeneity). After all, the primary goal is to estimate population parameters, and one cannot let an influential case or two unduly influence the estimates. In contrast, the methods shown in this paper represent a return to the person or persons in psychology. Because these methods are primarily visual in nature and do not rely on the computation of parametric statistics, outliers or assumptions of normality, homogeneity, etc., are never a concern. The Percent Correct Classification index is a simple frequency, and therefore an aggregate statistic, but it is always interpreted in light of a pattern (e.g., the *a priori* pattern or a multigram) and the complete set of observations. The simple "eye test" or more severe "interocular traumatic test" (Edwards et al., 1963) is taken seriously in OOM as there simply is no substitute for examining the data, particularly in light of an integrated model.

The final move, then, is from variable-based models to persons. The example study above employed a between-group design, and only two pertinent observations were made for each participant. A more intensive study of the individual is possible, however, using similar methods to conceptualize and analyze multiple observations made for each person. Cohn et al. (2014) for example, collected daily diary ratings from 54 women who had been raped. Ratings of post-traumatic stress disorder (PTSD) symptoms, drinking behavior, emotional states, and many other attributes, attitudes or behaviors were collected for 14 consecutive days. Using the OOM software in a novel analysis of the data, Grice et al. (in press) were able to examine a mediation model (PTSD → Negative Affect → Alcohol Consumption) at the level of the individual women. Unlike the aggregate results obtained from a variable-based Hierarchical Linear Model, the OOM analyses identified the individual women whose observations revealed a causal connection for each link in the model. In the world of clinical intervention where individuals—not means—are treated, such techniques are tantamount (Mumma, 2004; Haynes et al., 2009). Additional examples of person-centered studies using OOM have been published (e.g., Brown and Grice, 2012; Craig et al., 2012; Abramson et al., 2015), and methods of data analysis which permit a dynamic study of individuals have also been developed (e.g., see Valsiner et al., 2014). The time is therefore ripe for psychologists to return to a study of the person as an integrated, individual whole.

References

Abramson, C. I., Craig, D. P. A., Varnon, C. A., and Wells, H. (2015). The effect of ethanol on reversal learning in honey bees (*Apis mellifera* anatolica): response inhibition in a social insect model. *Alcohol* 49, 245–258. doi: 10.1016/j.alcohol.2015.02.005

Ayduk, O., May, D., Downey, G., and Higgins, E. (2003). Tactical differences in coping with rejection sensitivity: the role of prevention pride.

Pers. Soc. Psychol. Bull. 29, 435–448. doi: 10.1177/0146167202250911

Bakan, D. (1954). A generalization of Sidman's results on group and individual functions, and a criterion. *Psychol. Bull.* 51, 63–64. doi: 10.1037/h0058163

Behrens, J. T., and Yu, C.-H. (2003). "Exploratory data analysis," in *Handbook of Psychology*, Vol. 2, eds J. A. Schinka and W. F. Velicer (New York: Wiley), 33–64.

Berk, R. A., and Freedman, D. A. (2003). "Statistical assumptions as empirical commitments," in *Law, Punishment, and Social Control: Essays in Honor of Sheldon Messinger*, 2nd Edn, eds T. G. Blomberg and S. Cohen (New York: Aldine de Gruyter), 235–254.

Brown, E. A., and Grice, J. W. (2012). Mediation analysis via observation oriented modeling. *Int. J. Sci.* 1, 1–42.

Carlson, R. (1971). Where is the person in personality research? *Psychol. Bull.* 75, 203–219. doi: 10.1037/h0030469

Cohen, J. (1994). The earth is round ($p < 0.05$). *Am. Psychol.* 49, 997–1003. doi: 10.1037/0003-066X.49.12.997

Cohn, A., Hagman, B. T., Moore, K., Mitchell, J., and Ehlke, S. (2014). Does negative affect mediate the relationship between daily PTSD symptoms and daily alcohol involvement in female rape victims? Evidence from 14 days of interactive voice response assessment. *Psychol. Addict. Behav.* 28, 114–126. doi: 10.1037/a0035725

Craig, D. P. A., Grice, J. W., Varnon, C. A., Gibson, B., Sokolowski, M. B. C., and Abramson, C. I. (2012). Social reinforcement delays in free-flying honey bees (*Apis mellifera* L.). *PLoS ONE* 7:e46729. doi: 10.1371/journal.pone.0046729

Danziger, K. (1990). *Constructing the Subject*. Cambridge, UK: Cambridge University Press.

Downey, G., and Feldman, S. I. (1996). Implications of rejection sensitivity for intimate relationships. *J. Pers. Soc. Psychol.* 70, 1327–1343. doi: 10.1037/0022-3514.70.6.1327

Edwards, W., Lindman, H., and Savage, L. (1963). Bayesian statistical inference for psychological research. *Psychol. Rev.* 70, 193–242. doi: 10.1037/h0044139

Flórez, J. A. (2014). Peirce's theory of the origin of abduction in Aristotle. *Trans. Charles S. Peirce Soc. Q. J. Am. Philos.* 50, 265–280. doi: 10.2979/trancharpeirsoc.50.2.265

Grice, J. W. (2011). *Observation Oriented Modeling: Analysis of Cause in the Behavioral Sciences*. New York, NY: Academic Press.

Grice, J. W. (2014). Observation oriented modeling: preparing students for the research in the 21st Century. *Innov. Teach.* 3, 3.

Grice, J. W., Barrett, P. T., Schlimgen, L. A., and Abramson, C. I. (2012). Toward a brighter future for psychology as an observation oriented science. *Behav. Sci.* 2, 1–22. doi: 10.3390/bs2010001

Grice, J. W., Cohn, A., Ramsey, R. R., and Chaney, J. M. (in press). On muddled reasoning and mediation modeling. *Basic Appl. Soc. Psychol.*

Grice, J. W., Jackson, B. J., and McDaniel, B. L. (2006). Bridging the idiographic-nomothetic divide: a follow-up study. *J. Pers.* 74, 1191–1218. doi: 10.1111/j.1467-6494.2006.00407.x

Haig, B. D. (2005). An abductive theory of scientic method. *Psychol. Methods* 10, 371–388. doi: 10.1037/1082-989X.10.4.371

Haig, B. D. (2014). *Investigating The Psychological World*. Cambridge, MA: MIT Press.

Haynes, S. N., Mumma, G. H., and Pinson, C. (2009). Idiographic assessment: conceptual and psychometric foundations of individualized behavioral assessment. *Clin. Psychol. Rev.* 29, 179–191. doi: 10.1016/j.cpr.2008.12.003

Heathcote, A., Brown, S., and Mewhort, D. (2000). The power law repealed: the case for an exponential law of practice. *Psychol. Bull. Rev.* 7, 185–207. doi: 10.3758/BF03212979

Lambdin, C. (2011). Significance tests as sorcery: science is empirical—significance tests are not. *Theory Psychol.* 22, 67–90. doi: 10.1177/0959354311429854

Lamiell, J. T. (1997). "Individuals and the differences between them," in *Handbook of Personality Psychology* eds R. Hogan, J. Johnson, and S. Briggs (San Diego: Academic Press), 118–141.

Lamiell, J. T. (2003). *Beyond Individual and Group Differences: Human Individuality, Scientific Psychology, and William Stern's Critical Personalism*. Thousand Oaks, CA: Sage Publications.

Lamiell, J. T. (2013). Statistics in personality psychologists' use of trait constructs: what is it? How was it contracted? Is there a cure? *New Ideas Psychol.* 31, 65–71. doi: 10.1016/j.newideapsych.2011.02.009

Lebel, E. P., and Peters, K. R. (2011). Fearing the future of empirical psychology: bem's evidence of psi as a case study of deficiencies in modal research practice. *Rev. Gen. Psychol.* 15, 371–379. doi: 10.1037/a0025172

Manicas, P. (2006). *A Realist Philosophy of Social Science*. Cambridge, UK: Cambridge University Press.

Meehl, P. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *J. Consult. Clin. Psychol.* 46, 806–834. doi: 10.1037/0022-006X.46.4.806

Michell, J. (1999). *Measurement in Psychology: Critical History of a Methodological Concept*. Cambridge, UK: Cambridge University Press.

Michell, J. (2008). Is psychometrics pathological science? *Measurement* 6, 7–24. doi: 10.1080/15366360802035489

Michell, J. (2011). Qualitative research meets the ghost of Pythagoras. *Theory Psychol.* 21, 241–259. doi: 10.1177/0959354310391351

Molenaar, P., and Campbell, C. (2009). The new person-specific paradigm in psychology. *Curr. Dir. Psychol. Sci.* 18, 112–117. doi: 10.1111/j.1467-8721.2009.01619.x

Mumma, G. H. (2004). Validation of idiosyncratic cognitive schema in cognitive case formulations: an intraindividual idiographic approach. *Psychol. Assess.* 16, 211–230. doi: 10.1037/1040-3590.16.3.211

Rychlak, J. (1988). *The Psychology of Rigorous Humanism*, 2nd Edn. New York, NY: University Press.

Sidman, M. (1952). A note on functional relations obtained from group data. *Psychol. Bull.* 49, 263–269. doi: 10.1037/h0063643

Skinner, B. F. (1956). A case history in scientific method. *Am. Psychol.* 11, 221–233. doi: 10.1037/h0047662

Speelman, C. P., and McGann, M. (2013). How mean is the mean? *Front. Psychol.* 4:451. doi: 10.3389/fpsyg.2013.00451

Stevens, S. S. (1946). On the theory of scales of measurement. *Science* 103, 677–680. doi: 10.1126/science.103.2684.677

Thorngate, W. (1986). "The production, detection, and explanation of behavioral patterns," in *The Individual Subject and Scientific Psychology,* ed. J. Valsiner (New York: Plenum Press), 71–93.

Tukey, J. W. (1977). *Exploratory Data Analysis*. New York, NY: Pearson.

Valsiner, J. (1986). *The Individual Subject and Scientific Psychology*. New York: Plenum Press.

Valsiner, J., Molenaar, P., Lyra, M., and Chaudhary, N. (2014). *Dynamic Process Methodology in the Social and Developmental Sciences*. New York: Springer.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Grice. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

HYPOTHESIS AND THEORY
published: 09 May 2016
doi: 10.3389/fpsyg.2016.00674

Sampling Participants' Experience in Laboratory Experiments: Complementary Challenges for More Complete Data Collection

Alan McAuliffe and Marek McGann*

Department of Psychology, Mary Immaculate College, University of Limerick, Limerick, Ireland

Speelman and McGann's (2013) examination of the uncritical way in which the mean is often used in psychological research raises questions both about the average's reliability and its validity. In the present paper, we argue that interrogating the validity of the mean involves, amongst other things, a better understanding of the person's experiences, the meaning of their actions, at the time that the behavior of interest is carried out. Recently emerging approaches within Psychology and Cognitive Science have argued strongly that experience should play a more central role in our examination of behavioral data, but the relationship between experience and behavior remains very poorly understood. We outline some of the history of the science on this fraught relationship, as well as arguing that contemporary methods for studying experience fall into one of two categories. "Wide" approaches tend to incorporate naturalistic behavior settings, but sacrifice accuracy and reliability in behavioral measurement. "Narrow" approaches maintain controlled measurement of behavior, but involve too specific a sampling of experience, which obscures crucial temporal characteristics. We therefore argue for a novel, mid-range sampling technique, that extends Hurlburt's descriptive experience sampling, and adapts it for the controlled setting of the laboratory. This controlled descriptive experience sampling may be an appropriate tool to help calibrate both the mean and the meaning of an experimental situation with one another.

Keywords: averages, qualitative methods, mixed-methods, phenomenology, validity

OPEN ACCESS

Edited by:
Pietro Cipresso, Istituto di Ricovero e Cura a Carattere Scientifico – Istituto Auxologico Italiano, Italy

Reviewed by:
Leonard Bliss, Florida International University, USA
M. Teresa Anguera, University of Barcelona, Spain
Benjamin P. Chapman, University of Rochester Medical Center, USA

****Correspondence:***
Alan McAuliffe
alan.mcauliffe@mic.ul.ie

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *30 November 2015*
Accepted: *22 April 2016*
Published: *09 May 2016*

Citation:
McAuliffe A and McGann M (2016) Sampling Participants' Experience in Laboratory Experiments: Complementary Challenges for More Complete Data Collection. Front. Psychol. 7:674. doi: 10.3389/fpsyg.2016.00674

INTRODUCTION: TWO COMPLEMENTARY CHALLENGES

It is something of a trite observation amongst psychologists that not everything that matters can be measured. While a truism, any good psychologist also takes this as a challenge. We are aware, sometimes painfully so, of the limitations of our methods, and the complexity of our subject matter. But good science uses a range of techniques that complement one another and allows us to piece together a multiplex but increasingly coherent understanding of the mind and behavior. While some things cannot be measured, they can be observed and analyzed in rigorous and systematic ways that acknowledge and work within the boundaries of valuable data collection.

Our statistics are part of this toolbox of various methods that we use to build an understanding of psychology. Speelman and McGann (2013) reviewed a number of limitations of the mean as a representation of varied measurements, and the kinds of research designs built around their analysis. Their aim in doing so was not to be pessimistic about the possibility of accurate or valid

measurement in psychological science, but to prompt a discussion on the ways in which means or averages have been used uncritically and how their use might be improved as part of a wider effort to sharpen research practices in the discipline.

Speelman and McGann (2013) suggest no single means of improving care or practice with regards to the mean. Rather, a critical attitude that keeps theoretical assumptions in sight and reinforces an awareness of the derived nature of the mean (as opposed to it being assumed a measurement of an underlying parameter) is suggested. Mathematical and methodological techniques help refine the reliability of averages, helping to improve our confidence that an average indicates something important and stable about the data that have been collected. But we must also use varied methodological techniques to critically examine the *validity* of those data.

Speelman and McGann (2013) identify a number of assumptions in play in common use of the mean to summarize performance by an individual or group on a given task. The mean is typically used as an estimate of a "true" value being measured, with variability around that mean being a result of noise or other independent variables unrelated to those addressed in the experiment at hand. There are surely many cases where these assumptions hold true, but Speelman and McGann (2013) note that we should also be prepared to test these assumptions as a matter of common good practice.

We should be sensitive to the possibility that variability around the mean may have something important to tell us about the value of that statistic, and we are in need of techniques that allow us to interrogate such variations. Paying attention to variation in task performance could potentially enable us to validate our measurements, reinforce our interpretations, while also giving us a chance to spot new relevant variables, or other forms of confound.

Part of these efforts after validity involves the use of varied data gathering techniques, making a range of observations that might allow new information to come to the fore, and providing insights into patterns of behavior that might otherwise go unnoticed.

Each variable noticed can potentially be isolated, measured, and its contribution to a given set of performances teased apart through experimental or statistical control – in essence refining the mean being measured, distilling out the particular variable of interest from a complex mixture. There are some variables that have proven very difficult to quantify, isolate, and control, despite there being clear evidence that they play a role in how a person reacts to the task, materials, or situation of our laboratory experiments. In particular, the experience of the situation for participants, what the task or actions involved mean for them as they carry out the task, is something that tends to see little systematic analysis in experimental research, but has been increasingly recognized in recent years (Barrett et al., 2010, 2011). In the rest of the current paper we outline some *prima facie* reasons why a participant's experience of the laboratory and the apparent meaning of the task for them should be taken seriously. We then review some of the reasons, both historical and scientific, why the systematic collection of data concerning participants' experience remains relatively rare.

We thus outline two challenges that we suggest are somewhat complementary. On the one hand, the use of the mean in empirical studies demands a set of practices that police its validity. On the other, understanding the meaning of a situation requires the collection of remarkably difficult data – experiential reports – that are quintessentially un-averageable. If we are to test and refine the validity of our data, we will need to be able to find some way of examining variation in measured performance that might fit or diverge from variation in observed experiences. We review a number of different techniques for collecting experiential data and argue that, while useful in their current form, could yet be refined to provide us a more effective means of validating and calibrating measurements in laboratory behavioral experiments. While mixed methods approaches are becoming increasingly prevalent (Tashakkori and Teddlie, 2010), and have been deployed in a wide range of settings (REFS), we suggest that there remains a need for a new form of research method that more closely allies standard laboratory experiments with the collection of reports of participants' specific experiences of those experiments.

VALIDITY, EXPERIENCE, AND EXPERIMENTAL CONTROL

Assessing the validity of our measures is made difficult by the fact that it cannot be achieved via a single method. Though we might have a perfectly reliable measure, certainty regarding what it is that we are actually measuring comes not from the consistency of its numbers, but from our understanding of the tool and the ways in which it is used. The understanding that is vital to validity comes from approaching the same phenomenon from other angles, using other methods. No measurement is pure and no experiment perfect, but over time and through the convergence of multiple points of view we gradually develop a picture of our subject matter in increasingly fine resolution. Where validity of the mean, in particular, is concerned, we will need several complementary studies of a behavioral phenomenon that make it clear it is reliable, and insofar as the meaning or experience of the situation is one of the things that cause it to vary, that we sample those as appropriate.

Decades of research in Psychology have taught us that in the experiments where we make our measurements, meaning matters a great deal. Meaning has been on the agenda in some form or another since the "New Look" studies of Bruner and colleagues, which played a substantial role in the rise of cognitive psychology. Bruner and Goodman (1947) reported that coins were perceived or remembered as having different sizes depending on the economic status of the person doing the perceiving, while Bruner and Postman (1949) showed error and expectancy effects due to prior experience and understanding of decks of playing cards. Bruner (1990) has since distanced himself from the computationalist understanding of the mind that developed in part from this line of work on perception, but maintains that understanding the role of meaning in psychology is vital if we are to advance the science, advancing a theory of

meaning as culturally enacted but still constitutive of cognitive activity.

The classic work of Treisman (1960), still cited in introductory texts to cognitive psychology, illustrated how people's attention often moves fluidly with the meaning of the stimuli they are being exposed to, rather than the particular sensory channel on which they were supposed to be focusing. While such research as the New Look and experiments on attention made it clear that the meaning of the stimuli matter for so-called "lower level" aspects of cognition, decades of research were triggered when Wason (1971) showed that it affects reasoning too. People reason to different inferences depending on whether the material they were working with were meaningful to them – whether the materials fit a person's general experience of the world – or whether they were abstract and contrived.

Perhaps more pointedly, research on participants' experience of psychological research itself highlights the potency of a situation's apparent meaning for people's behavior. Since Orne's (1962) exploration of demand characteristics, we have been sensitive to the fact that participants who interpret the experiment as testing a particular hypothesis tend to skew their behavior (either deliberately or unconsciously) to support or undermine the perceived hypothesis. Orne (1973) argued that people respond to the "total experimental situation" and that a range of steps should be taken to cope with the rather holistic nature of the setting influencing people. Orne's work itself developed within a context of increasing disciplinary recognition that the stimulus materials were only part of the picture in understanding behavior in psychology experiments.

Rosenberg (1969) reported three conditions of a study in which participants were asked how much they liked or disliked various pictured persons. Both groups were informed that past research indicated that liking–disliking reactions to strangers correlated with maturity. One group were told that psychologically mature and healthy individuals show greater liking for strangers than immature people and were given fabricated journal article citations. The other experimental group were told the opposite – that research indicated that immaturity was associated with greater liking of strangers, with fabricated journal articles cited. Both groups, however, were informed that they were not going to take part in a study of liking–disliking images of strangers, but rate pictures of strangers to create a standardized list of photographs. Participants believed that these photographs were then going to be used in a liking–disliking task in future research. It isn't surprising that there were significant differences between the groups, but the obvious manipulation here is not the full story. Rosenberg's work is a clear illustration of evaluation apprehension, which can be made to affect experimental responding. However, Rosenberg also included a control group with no information about maturity and liking. The results indicated that male participants in this neutral context condition rated male pictures much lower than both experimental groups. They even rated the images substantially lower than the group that were informed that lower ratings was associated with maturity.

Expectancy, social desirability, and demand effects within psychological research are all indications that what participants are doing is not naively fixed by the explicit instructions presented to them, but richly enmeshed with the meaning of the context as a whole. The average response to a given task or stimulus is a product not of a single fix instruction set, but a varied participant-lab situation.

More subtly, work by Gallagher and Marcel (1999) with patients with dyspraxia indicates how their performance on a given task varies substantially with its meaningfulness. Very similar bodily movements that are difficult or impossible for a patient in clinical assessment might be performed relatively smoothly and effectively in situations where the context is more meaningful for them. Lifting a cylindrical object from a table might be a challenge, but taking a drink of water from a tumbler straightforward. Touching their nose on demand can be difficult, but pushing their glasses back into position is done without pause for thought.

More recently we have seen a renewed surge in interest in context, and how it is defined not just by the stipulations of the experimenter but by the total situation involving the thoughts, feelings and behaviors of a particular person, at a particular time (Barrett et al., 2010, 2011; Schwarz, 2010). The experience of the participant and the meaning of the situation for them is once again being acknowledged and given a central role in how we consider their behavior. If we are to adequately understand what a person does, so the understanding goes, we cannot just examine the "input", the stimuli used, the wording of instructions, or the logical details of the task in which the person was engaged. The validity of our measures is derived from the whole situation and should be examined within the context of that whole situation – including their own experience of it. Though there is no claim that this is *all* that matters, this is one facet of the complexity of a laboratory situation affecting the value and variability of measurements made in that situation, and which should be included as a consideration when policing the validity of those measurements across replications.

Several related threads of theoretical and empirical work share this concern with experience. They tend to vary, however, in terms of their descriptions of the relationships between experience and behavior (Thompson, 2007; Di Paolo, 2009; Shapiro, 2010; Wilson and Golonka, 2013) though most commonly the specifics of that relationship remain ill-defined.

There are thus long threads of research through the history of experimental psychology, including many that have become increasingly influential in recent years, that make a strong case for including some account of the participants' experience of the experiment in our analyses and interpretation of the data (or at least some aspects of the data). Swinging against this trend, however, is one with an even longer history within the discipline pointing to the weaknesses and unreliability of people's description of their own thoughts and behavior.

Good Reasons to Distrust Experiential Reports

While it is clear that people's experience matters to their behavior, more than a century of research has shown us that it is difficult to understand just *how* it matters. Scientific psychology had

the examination of consciousness at its core during the period when all of its major institutions were founded. However, several decades of the analysis of experience ground to a halt in the face of difficulties with introspection. The difficulties of shared analysis, the challenges of independent testing, and the existence of unfalsifiable claims, all made consciousness a problematic notion for a burgeoning science (Watson, 1913; Fancher, 1996; Richards, 2002).

Experience was marginalized by most forms of behavioristic psychology that dominated research through the middle half of the twentieth century. When interest arose again in latter decades, much of the research showed that what effects the meaning of a situation might have for participants' behavior, can occur without them being consciously aware of it. As such, people are poor describers of their own behavior, or the reasons for it. Perhaps most famously, Nisbett and Wilson's (1977) review supporting the idea that people have little to no insight into the causes and influences on their own behavior drove home just how poor a source of data individual's self-report is when we are interested in understanding their actions. Not only does it seem that we do not accurately experience the causes of our actions, but we are happy to invent reasons or explanations that bear little relation to what those real influences are.

Johansson et al.'s (2005) instant classic work on "choice blindness" more recently illustrated just how quickly we can produce such confabulations. Participants, when asked to choose the more attractive between two photographs, and then asked to explain their decision after being handed the *wrong* photo still offered reasons, some mentioning unique aspects of the new (unchosen) picture. Later work showed these confabulated justifications for events to be insensitive to what actually happened (Johansson et al., 2006).

Relatedly, Marcel's (1993) work on multiple modes of response indicates that we can simultaneously be conscious of a stimulus in one response modality but not in another. That is, if asked to speak a response or press a button, the same stimulus might be simultaneously in a person's experience and not. Experience, whatever it might be, cannot be understood as a single, simple stream of thought tightly bound to our behavior (Dennett, 1991).

Work in the neuroscience of vision seems to compound this distinction between experience and action through the identification of two apparently quite separate streams of visual processing in the brain (Milner and Goodale, 1995; Goodale and Milner, 2005). One, the dorsal stream, seems specialized for the coordination of visuo-motor action, enabling a person to engage effectively with objects through visual cues. The other, ventral stream, appears to process the visual awareness of objects, dealing with object recognition and naming. Various forms of so-called "blindsight" illustrate the dissociation between these two streams, where a person's experience can partially or dramatically disrupted while their actions remain effective (Milner and Goodale, 1995).

The consistent trend throughout research on consciousness and behavior is that the linkage between these two aspects of psychology is not straightforward. Understanding that relationship will not come from any casual introspection or direct insight from people reporting what they think. In the existing research the tendency is to explore people's awareness of their own actions, the reasons for those actions, or in the case of the likes of Marcel's work, their responses to minimally relevant stimuli – that is images or sounds that only matter to the participant within the constraints of the research task. To that extent the research has tended to focus either on a person's already conceptualized, considered experience – their *metacognitive* awareness of their thought and actions – or on tasks that are stripped of meaningful context for people and therefore do not fit easily within their normal range of behavior or their normal experiences.

The recent rise in interest concerning context, experience and meaning noted above (see e.g., Varela et al., 1991; Lutz, 2007; Barrett et al., 2010; Mesquita, 2010; Schwarz, 2010; Froese et al., 2011a,b) has criticized such pre-interpreted data. While we must clearly be wary of the claims about their experience and their behavior that we elicit from our participants, there might still be important information we should collect from them about the experience itself. These recent trends lean toward including the analysis of some form of "raw" experience in the interpretation of behavioral data, and perhaps the interrogation of variability within those data. The existing research makes it clear that there is a strong relationship between the participant's experience, what the situation means to them, and their behavior. It is equally clear that this relationship, however, strong, is complicated. There is no tight coupling between how a person experiences a situation or stimulus, and the fine-grained details of their behaviors in response.

That the existing research leaves us in such a state of confusion suggests that the manner in which we have been collecting data concerning experience is limited, and that other methods are required. We must be careful and nuanced in our gathering and interpreting of experiential reports. While people may provide poor explanations for their actions, their reports of just what they experienced may nevertheless hold valuable information for psychological researchers. Over the past two decades a number of different research methods have developed that may improve matters. We argue that while these methods certainly advance the science of the relationship between experience and action, and can therefore help explore some of the issues regarding variability in behavior on the basis of the meaning of the laboratory situation for the participant, there remains room for refinement.

NEWER TECHNIQUES FOR THE STUDY OF EXPERIENCE: WIDE AND NARROW APPROACHES

Different approaches to studying experience come with different commitments to levels of analysis, timescales of measurement, and quality of information regarding the person's activity at the time of the experience being examined. Some methods, which we will here term 'wide' approaches to experience, gather reports or observations in a manner that involves less structure or deliberation with regards to the activity in which the person is engaged at the time, but tends to maximize the range of possible responses and is often captured in ecologically relevant activities.

Examples of such wide approaches are most standard qualitative research methods in psychology, such as interviewing or focus groups (Banister, 2011), the bottom-up explorations of interpretative phenomenological analysis (Reid et al., 2005; Palmer et al., 2010), and descriptive experience sampling (DES; Hurlburt and Heavey, 2002; Hurlburt et al., 2002; Hurlburt and Akhter, 2006), with its randomized triggering of introspective episodes.

Wide approaches gather less constrained information, and in doing so enable a broader exploration of possible research questions. While it is possible to explore the relationship between experience and actions with these methods, this tends to produce a high level, low-resolution picture. These kinds of analyses are useful in pointing us in the direction of more specific research questions, and identifying broader patterns that are difficult if not impossible to see using more narrowly focused methods.

Interviewing and focus groups, for instance, allow us to explore people's concepts of what they are doing, or how they understand the situation in which they find themselves (Banister, 2011). When participants' understanding is our key point of interest, this is valid. However, where our interest is in understanding the specifics of the relationship between actions and behavior, things break down, as the classic work on this issue in experimental studies has shown.

Intepretative phenomenological analysis (IPA) is modest in its aims in that it eschews claims to produce facts or unbiased data, but notes that most people are not naive in their experiences – they are experts, or at least familiar with the kinds of situations in which they typically find themselves (Reid et al., 2005). In partnership with a researcher people can reflect upon and interpret their experience using all of the richness of history and context that they bring to the situation, enabling the exploration of certain kinds of relationships unavailable to many more mainstream research techniques. The data typically collected for IPA are interview transcripts, and as such depend on the participants' recollection for the event or events being examined. Where the particular coupling of experience and behavior is of interest, there are quite strong limits on what kind of insights this form of analysis will enable.

Descriptive experience sampling aims to access "pristine" (Hurlburt and Akhter, 2006) experience, relying less on retrospective accounts of an experience, more on notes and recorded comments made in the moments immediately following an instant of experience, prompted by a beeper device or similar trigger. The pristine nature of the experience – that it is within the flow of the person's natural activity, sampled without much warning by a randomly occurring trigger – is at the heart of the method's intended use. Random sampling, and the uncontrolled character of the environment mean that the possibility of associating experiences with particular behaviors is once again limited (though not entirely ruled out, see Hurlburt et al., 2002).

Wide approaches to the study of experience are open to the flow of experience and behavior within naturally occurring activity. In approaches that are both qualitative and mixed methods, these techniques have been applied in domains such as Nursing (e.g., Traylen, unpublished MPhil dissertation), Education (e.g., Onwuegbuzie et al., 2007; Palak and Walls, 2009), Anthropology (e.g., Killick, 1998), as well as Psychology (e.g., Hurlburt and Akhter, 2006). They offer useful insights into the relationship between experience and behavior, and can be used to help structure sequential mixed methods research projects where concepts and experiences are sampled in ecologically rich settings and then variables identified for closer inspection in laboratory experiments. For the more fine-grained examination of specific variability of behavior in those experiments, however, these approaches tend to be too broad, examining timescales that are too long to adequately sample experience at the grain of analysis that the behavior is being measured.

"Narrow" approaches, on the other hand, focus more particularly at the level of momentary experience and momentary behaviors. In a sense, the entire domain of psychophysics exists at this level of analysis, a very longstanding and finely tuned examination of the relationship between physical stimuli and a person's experience of them. A somewhat related but distinct precedent in the methodological literature is that of systematic observation (Hintze et al., 2002; Podsakoff et al., 2003). Systematic observation, with a long history in various disciplines, clearly specifies the behaviors of interest in advance and observes them (and only them) in naturalistic settings. It therefore constitutes as more focused form of observation than the "wide" approaches outlined above. The technique tends not to involve the sampling of participants' experience or awareness of their surroundings at the moment of interest, however, and the measurements of behavior while specific, are typically more coarsely grained than would be common in controlled experiments (though this may change as technology advances).

In the present paper, our interest is specifically with the experience-behavior relationship, and how variability in experience might be used to better understand variability in measured behaviors. For that purpose we find two candidate approaches in recently developed methods for fine-grained experiential data collection: neurophenomenology (Varela, 1999; Lutz and Thompson, 2003; Thompson et al., 2005) and the elicitation interview (Petitmengin, 2006; Petitmengin et al., 2013).

Both neurophenomenology and the elicitation interview involve quite substantial control over the environment in which that data are collected. In the case of neurophenomenology the research is conducted in a neuroscience laboratory, usually with EEG recording, and involves the careful training of participants in phenomenological introspective techniques (that is, introspection that attempts to avoid conceptualisation of the experience, but to review and report it in as close to an atheoretical fashion as possible). Neurophenomenology is thus an example of a mixed methods approach (Tashakkori and Teddlie, 2010; Creswell and Plano Clark, 2011), seeking calibration of quantitative measures with qualitative reports. The elicitation interview is similarly conducted in a controlled setting, but in this case the participant is not trained to introspect but interviewed by a specialist in a manner intended to evoke the experience of a particular moment, as opposed to some particular *post hoc* understanding of that moment.

Being lab-based, both neurophenomenology and the elicitation interview offer the possibility of linking experience with reliably, and finely, measured behaviors. They provide the possibility of a high resolution examination of the relationship between experience and action. They are not, of course, without their drawbacks.

Neurophenomenology requires training of participants in the particular introspective techniques associated, and in doing so alters the very experience we are studying. Lutz and Thompson (2003) argue that this is not a deep problem, though do not offer a full explanation as to why. While it is quite possibly true that coming to an understanding of experience will necessarily change it, we would argue that methods should still be explored that might possibly provide us with naive or unreflective experiential reports. We do not argue against neurophenomenology, but simply note that there may yet be useful experiential data to collect from participants whose reports are not pre-disciplined by the training they have received. Neurophenomenology is one tool available to us, we note that others are yet needed.

The elicitation interview purports to provide just such naive data, and in this we see real promise, but two facets of the technique imply limits that might still leave us with an important methodological blindspot.

The collaboratively constructed nature of the interview process is one point of consideration, keenly aware as we already are about the ease with which apparently confabulated responses about experiential reports are produced. While proponents of the elicitation interview approach argue strongly that a properly skilled interviewer neither foists particular descriptions nor prompts invented reports from their interviewees (Petitmengin, 2006), we must yet proceed with care. This means that the approach, while both demanding of extraordinary discipline on the part of the interviewer and substantial time for its conduct (often between half an hour to an hour per interview), must still be used with caution. Such pragmatic considerations must not stop us from doing good science, but they do, nevertheless, motivate us to be fully cognisant of the range of choices we have available.

More concerning for our current purposes is the standard focus of the elicitation interview: the re-evocation of a particular moment of experience, an instant, as it were, during which a decision was made, or a response to a question as it popped into the interviewee's mind. The techniques of the interview bring the participant back to that moment, as though it were as real and rich as their immediate environment. With the previous experience thus being relived, it can be interrogated in fine detail. In doing so, however, the temporal relationship between event and subsequent discussion is broken. In Petitmengin et al.'s (2013) recent study on the Johansson et al. 's (2005, 2006) choice blindness task, for instance, some participants completed the photo choice and explanation at the normal pace, with reports on the decision occurring between 5 s and 1 min after the choice. The elicitation interview involved a period of between 30 and 45 min post-decision before re-presentation of the photo and evoking of explanation. It is very likely that the collection of systematic experiential reports of any kind is going to involve the interruption of the flow of behavior within a task in some form. We would argue, however, that more modest interruptions should be more attractive, and where possible the temporal dimensions of the task should be carefully balanced across participant groups. What is more important, however, is the possibility of multiple sampling points throughout the course of a task. Where highly focused techniques such as the elicitation interview provide fine-grained examination of a single moment, there is not only a possibility but some suggestive evidence of multiple strands of experience, and multiple rhythms of attention or endogenous sensitivity to different aspects of the environment operating over different timescales (Varela et al., 1981; Donald, 2001; Busch et al., 2009). That is, our experience is not just a string of beads, but has multiple tempos and currents to it that will need multiple sampling to observe, a form of repeated probing that the likes of the elicitation interview makes unfeasible.

We therefore argue that there is room between the wide and narrow forms of investigation of experience for a set of intermediate methods. This intermediate range is more anchored in recorded events and actions than wide approaches. Such an approach will enable it to be used within controlled environments, and thus offers promise in collecting data relevant to the interrogation of variable behavior in controlled settings. The approach would also, though, be less finely coupled to particular stimuli or instants of experience than the more narrow approaches. The meaningfulness of actions is to be sampled at this intermediate range, where we might find patterns of behavior rather than individual events, and themes of experience rather than fine-grained particulars. Instead of the fast, very short durations of most neural events as measured and used in neurophenomenology, we might explore the slower, 10s of seconds or minutes of duration in common behavior settings. Given the history of research on experience-behavior links, we might expect relationships between sampled experience and behavior to need this kind of re-sampling, so that variability in behavior can be calibrated against variability in experience, rather than trying to capture something fixed in either one.

SUGGESTING AN INTERMEDIATE LEVEL OF ANALYSIS

While dependencies of behavior on a host of contextual factors is violated in laboratory experiments, this is a compromise adopted for the purposes of maximizing communicability (through standardized meanings to terms and procedures) as well as replicability [an issue of some current concern amongst researchers (Koole and Lakens, 2012; Nosek et al., 2012; Open Science Collaboration, 2012, 2015; Ritchie et al., 2012; Roediger, 2012)].

Long running debates over the value of lab vs. field research are essentially the professional policing of this compromise, an exercise in maintaining perspective on the complementary values of different forms of data collection, and an effort at continually refining and improving our methods. The collection of reports of the experiences of participants is no exception to this issue, with wider approaches serving richer understandings of context, while the more narrowly focused techniques offer higher

resolution accounts of more finely circumscribed phenomena. Wide approaches explore the general attitudes and experiences of a person at a conceptual level that fits the person's understanding of their situation and actions, but that makes specific reference to particular experiences and behaviors challenging. Narrow approaches, on the other hand, may in fact be swamping the signal on the relationship between experience and behavior with the noise of momentary stream of consciousness, much of which is irrelevant to the niceties of bodily action (Aglioti et al., 1995; Milner and Goodale, 1995). If the meaning of the situation (as suggested by the likes of Barrett et al., 2010), rather than strings of isolated stimuli, are part of what matter to the structuring of behavior, and the variability of measurements around a mean for a given behavioral variable, then at least some of the varied methods we use should be calibrated at that appropriate scale.

Without knowing what experiential data most matter for best understanding behavior, the wise course of action is to sample widely and often, but within a setting where the behavior is sufficiently reliable to keep subtle relationships stable (or as stable as they can be). We suggest a form of controlled descriptive experience sampling (a "C-DES"), where introspective moments are triggered as with standard DES – without prior warning to the participant, via a beep or flash, perhaps. The participants might understand these triggers to be random, but they need not be in actuality. Descriptions can be kept brief, to potentiate multiple such sampling during a single task or event as appropriate. Further, the purely verbal descriptions of standard DES might also be augmented with simple video recording of non-verbal behaviors such as blinks, eye-movements, or other possibly subtle, aspects of the participant's behavior, offering a richer interpretative context for the content of reports (Olivares et al., 2015).

To offer an illustration, the Iowa Gambling Task (Bechara et al., 1994) is a frequently used laboratory activity conducted to evaluate participants' sensitivity to certain kinds of consequences, or to investigate trait characteristics such as impulsivity or executive control. The task is sometimes augmented with questions to the participant about their knowledge of its various components, to see how this changes over the course of the activity. Just what the relationship between participants' knowledge and their behavior is over the course of the task is somewhat problematic, but C-DES would eschew a need for the *participant* to understand the task at all, or report knowledge of it. Rather, by sampling what they were aware of either at key moments, or at regular intervals over the course of the task, researchers might be able to explore this relationship without relying on participant insight.

While this runs counter to the standard use of DES, for which naturalistic activity is vital, many of the strengths of the approach are maintained (no pre-specification or priming of behavior or moment to be introspected upon, naturalistic description of experience by participants). These strengths might thus be deployed in the service of understanding people's experiences of the laboratory during the laboratory task, and provide one of several perspectives from which we build up a richer understanding of what people are doing, and how they are experiencing the doing of it.

We will not know without conducting the research what kinds of experience will be relevant. History indicates clearly that introspective explanations of behavior are not the data we are looking for, but a plethora of other options are available, across numerous scales of time. Sensory experiences, physiological rhythms and responses, emotions, moods, culturally relevant routines – these things, and more show up in people's descriptions of their experience. While long-practiced habits might primarily shape behavior at the level of momentary particulars, experience may instead be coupled with action at the level of "molar behavior" (Barker, 1968).

This is to say that experience may not be a flow of individual moments in continuous accumulation, but a general awareness of a situation within which various relationships become distinguished – an event does not simply happen at some psychological "now", but early or late within a general expectation or understanding of the setting. Longstanding (but little known) work indicates that people are very sensitive to the standing patterns of behavior or expected routine present within a given physical or social setting (Barker, 1968; Schoggen, 1989; Heft, 2001, 2003, 2007; see also Heft et al., 2014, for a recent examination of people's ability to recognize settings with very limited information). The work of Mesquita (2010) and Barrett et al. (2011, 2014), have shown a similarly situational character to people's emotional reactions.

Within a more controlled form of DES the probing of conscious awareness can remain open and largely unstructured. Participants are free to describe their experience in familiar and comfortable terms, which can be explicated in conversation with the experimenter either immediately, or at a later time after the experimental task itself is completed. For the main, the standard DES principles outlined by Hurlburt et al. (2002) apply. The time between experience reporting and exploration in collaboration with the researcher is very short. Moments of experience are clearly defined (by the use of a tone or other trigger). Various practices of the interview are used to ensure that careful distinctions are made between the experience itself and any attempt to explain that experience.

In addition, however, given that the initial probings of experience can be kept brief (or varied in length depending on research goals), the possibility of multiple samplings over the course of a single experimental session is maintained. The intervals between samplings can be used as a means of exploring the temporal aspects of experience, its rhythms, and periodic variations.

USING THE UN-MEAN-ABLE TO CALIBRATION THE MEAN (AND VICE-VERSA)

Focusing closely on averages as summaries of collections of data is a practice that depends on a host of background theoretical assumptions. Speelman and McGann (2013) raised concerns (oft-noted in statistics courses, but rarely applied in practice) that

these assumptions are commonly unquestioned, and frequently ill-considered. While there are some reporting and analysis practices that might help contextualize the mean in mathematical or statistical terms, and we support calls to move toward standardizing such practices (such as Doherty et al., 2013), it is equally important to query the psychological, and not just statistical, context to the data being collected.

In this paper we have argued that there are good reasons for paying more attention than we typically do to the experience of the participant within rigorous laboratory experiments. There is clearly a relationship between participants' experiences of a given situation and their behavior within that situation, but the relationship is not a simple one. The validity of our measures, and relatedly our understanding of their variation, must be achieved through the coordination of multiple sources of knowledge about a person and their actions in a given setting. Experiential data, however, challenging they are to work with, have some role to play in that validation and calibration process (Froese et al., 2011a).

What we have termed "wide" approaches to such experiential data collection do not provide us with the behavioral data at the level of detail we need to effect this calibration. Conversely, the approaches we have termed "narrow" we suggest are *too* narrow. Though they enable the collection of specific behavioral data, the pre-focused nature of their experience sampling imposes expectations or prior understandings of the kinds of experience we need to probe, and include assumptions about the momentary nature of those experiences, that are inappropriate for our current levels of understanding (or perhaps more accurately, ignorance), about the behavior-experience relationship, particularly of the varying timescales of different phenomena of consciousness.

We propose that a C-DES is a data collection technique ideal for the kinds of disciplined exploratory research that is needed to adequately observe the experience-behavior relationship. In order to determine to what degree a calculated mean actually matters to what people do, and how to refine the validity of what it measures, we need a level of description and analysis of experiential data that is not commonly in use – one that is exploratory and potentially wide-ranging, but evoked within a controlled, managed situation such as the laboratory experiment. The paired examination of controlled behaviors still offers us a means of understanding and interpreting the descriptions of experiences captured through this process. The validation of the mean and the un-meanable is a two-way relationship, achieved not through a single ideal study, but through a long process of negotiation across multiple studies, using multiple methods.

AUTHOR CONTRIBUTIONS

All authors listed, have made substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENT

The work in this paper constitutes part of AA's doctoral studies, supervised by MG.

REFERENCES

Aglioti, S., De Souza, J. F., and Goodale, M. A. (1995). Size-contrast illusions deceive the eye but not the hand. *Curr. Biol.* 5, 679–685. doi: 10.1016/S0960-9822(95)00133-3

Banister, P. (2011). *Qualitative Methods in Psychology: A Research Guide*. London: McGraw-Hill.

Barker, R. G. (1968). *Ecological Psychology: Concepts and Methods for Studying the Environment of Human Behavior*. Stanford, CA: Stanford University Press.

Barrett, L. F., Mesquita, B., and Gendron, M. (2011). Context in emotion perception. *Curr. Dir. Psychol. Sci.* 20, 286–290. doi: 10.1177/0963721411422522

Barrett, L. F., Mesquita, B., and Smith, E. R. (2010). "The context principle," in *The Mind in Context,* 1st Edn, eds B. Mesquita, L. F. Barrett, and E. R. Smith (London: Guilford Press).

Barrett, L. F., Wilson-Mendenhall, C. D., and Barsalou, L. W. (2014). "The conceptual act theory: a roadmap," in *The Psychological Construction of Emotion*, eds L. F. Barrett and J. Russell (New York, NY: Guildford Press).

Bechara, A., Damasio, A. R., Damasio, H., and Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition* 50, 7–15. doi: 10.1016/0010-0277(94)90018-3

Bruner, J. (1990). *Acts of Meaning*. Cambridge, MA: Harvard University Press.

Bruner, J. S., and Goodman, C. C. (1947). Value and need as organizing factors in perception. *J. Abnorm. Soc. Psychol.* 42, 33–44. doi: 10.1037/h0058484

Bruner, J. S., and Postman, L. (1949). On the perception of incongruity: a paradigm. *J. Pers.* 18, 206–223. doi: 10.1111/j.1467-6494.1949.tb01241.x

Busch, N. A., Dubois, J., and VanRullen, R. (2009). The phase of ongoing EEG oscillations predicts visual perception. *J. Neurosci.* 29, 7869–7876. doi: 10.1523/JNEUROSCI.0113-09.2009

Creswell, J. W., and Plano Clark, V. L. (2011). *Designing and Conducting Mixed Methods Research*. London: Sage.

Dennett, D. C. (1991). *Consciousness Explained*. London: Penguin.

Di Paolo, E. (2009). Extended life. *Topoi* 28, 9–21. doi: 10.1007/s11245-008-9042-3

Doherty, M. E., Shemberg, K. M., Anderson, R. B., and Tweney, R. D. (2013). Exploring unexplained variation. *Theory Psychol.* 23, 81–97.

Donald, M. (2001). *A Mind So Rare*. London: Norton.

Fancher, R. (1996). *Pioneers of Psychology,* 3rd Edn. New York, NY: W. W. Norton & Co.

Froese, T., Gould, C., and Barrett, A. (2011a). Re-viewing from within: a commentary on first-and second-person methods in the science of consciousness. *Construct. Found.* 6, 254–269.

Froese, T., Gould, C., and Seth, A. K. (2011b). Validating and calibrating first-and second-person methods in the science of consciousness. *J. Conscious. Stud.* 18, 38–64.

Gallagher, S., and Marcel, A. J. (1999). The self in contextualized action. *J. Conscious. Stud.* 6, 4–30.

Goodale, M. A., and Milner, A. D. (2005). *Sight Unseen: An Exploration of Conscious and Unconscious Vision*. Oxford: Oxford University Press.

Heft, H. (2001). *Ecological Psychology in Context: James Gibson, Roger Barker, and the Legacy of William James's Radical Empiricism,* 1st Edn. London: Lawrence Erlbaum Associates.

Heft, H. (2003). Affordances, dynamic experience, and the challenge of reification. *Ecol. Psychol.* 15, 149–180. doi: 10.1207/S15326969ECO1502_4

Heft, H. (2007). The social constitution of perceiver-environment reciprocity. *Ecol. Psychol.* 19, 85–105. doi: 10.1080/10407410701331934

Heft, H., Hoch, J., Edmunds, T., and Weeks, J. (2014). Can the identity of a behavior setting be perceived through patterns of joint action? An investigation of place perception. *Behav. Sci.* 4, 371–393. doi: 10.3390/bs4040371

Hintze, J. M., Volpe, R. J., and Shapiro, E. S. (2002). "Best practices in the systematic direct observation of student behaviour," in *Best Practices in School Psychology*

IV Vol. 2, eds A. Thomas and J. Grimes (Bethesda, MD: National Association of School Psychologists), 993–1006.

Hurlburt, R. T., and Akhter, S. A. (2006). The descriptive experience sampling method. *Phenom. Cogn. Sci.* 5, 271–301. doi: 10.1007/s11097-006-9024-0

Hurlburt, R. T., and Heavey, C. L. (2002). Interobserver reliability of descriptive experience sampling. *Cogn. Ther. Res.* 26, 135–142. doi: 10.1023/A:1013849922756

Hurlburt, R. T., Koch, M., and Heavey, C. L. (2002). Descriptive experience sampling demonstrates the connection of thinking to externally observable behavior. *Cogn. Ther. Res.* 26, 117–134. doi: 10.1023/A:1013849922756

Johansson, P., Hall, L., Sikström, S., and Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science* 310, 116–119. doi: 10.1126/science.1111709

Johansson, P., Hall, L., Sikström, S., Tärning, B., and Lind, A. (2006). How something can be said about telling more than we can know: on choice blindness and introspection. *Conscious. Cogn.* 15, 673–692. doi: 10.1016/j.concog.2006.09.004

Killick, D. (1998). "On the value of mixed methods in studying mining communities," in *Social Approaches to an Industrial Past: The Archaeology and Anthropology of Mining*, eds A. B. Knapp, V. C. Piggott, and E. W. Herbert (New York, NY: Wiley), 279–290.

Koole, S. L., and Lakens, D. (2012). Rewarding replications a sure and simple way to improve psychological science. *Perspect. Psychol. Sci.* 7, 608–614. doi: 10.1177/1745691612462586

Lutz, A. (2007). Neurophenomenology and the study of self-consciousness. *Conscious. Cogn.* 16, 765–767. doi: 10.1016/j.concog.2007.08.007

Lutz, A., and Thompson, E. (2003). Neurophenomenology integrating subjective experience and brain dynamics in the neuroscience of consciousness. *J. Conscious. Stud.* 10, 31–52.

Marcel, A. J. (1993). "Slippage in the unity of consciousness," in *Experimental and Theoretical Studies of Consciousness*, eds G. R. Rock and J. Marsh (Chichester: John Wiley & Sons), 168–180.

Mesquita, B. (2010). "Emoting: a contextualized process," in *The Mind in Context,* 1st edn, eds B. Mesquita, L. F. Barrett, and E. R. Smith (London: Guilford Press).

Milner, D., and Goodale, M. A. (1995). *The Visual Brain in Action*. Oxford: Oxford University Press.

Nisbett, R. E., and Wilson, T. D. (1977). Telling more than we can know: verbal reports on mental processes. *Psychol. Rev.* 84, 231–259. doi: 10.1037/0033-295X.84.3.231

Nosek, B. A., Spies, J. R., and Motyl, M. (2012). Scientific utopia II. Restructuring incentives and practices to promote truth over publishability. *Perspect. Psychol. Sci.* 7, 615–631. doi: 10.1177/1745691612459058

Olivares, F. A., Vargas, E., Fuentes, C., Martínez-Pernía, D., and Canales-Johnson, A. (2015). Neurophenomenology revisited: second-person methods for the study of human consciousness. *Front. Psychol.* 6:673. doi: 10.3389/fpsyg.2015.00673

Onwuegbuzie, A. J., Witcher, A. E., Collins, K. M., Filer, J. D., Wiedmaier, C. D., and Moore, C. W. (2007). Students' perceptions of characteristics of effective college teachers: a validity study of a teaching evaluation form using a mixed-methods analysis. *Am. Educ. Res. J.* 44, 113–160. doi: 10.3102/0002831206298169

Open Science Collaboration (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspect. Psychol. Sci.* 7, 657–660. doi: 10.1177/1745691612462588

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science* 349, aac4716. doi: 10.1126/science.aac4716

Orne, M. T. (1962). On the social psychology of the psychological experiment: with particular reference to demand characteristics and their implications. *Am. Psychol.* 17, 776–783. doi: 10.1037/h0043424

Orne, M. T. (1973). "Communication by the total experimental situation: why it is important, how it is evaluated, and its significance for the ecological validity of findings," in *Communication and Affect: Language and Thought (pp. xii, 200)*, eds P. Pliner, L. Krames, and T. Alloway (New York, NY: Academic Press).

Palak, D., and Walls, R. T. (2009). Teachers' beliefs and technology practices: a mixed-methods approach. *J. Res. Technol. Educ.* 41, 417–441. doi: 10.1080/15391523.2009.10782537

Palmer, M., Larkin, M., de Visser, R., and Fadden, G. (2010). Developing an interpretative phenomenological approach to focus group data. *Qual. Res. Psychol.* 7, 99–121. doi: 10.1080/14780880802513194

Petitmengin, C. (2006). Describing one's subjective experience in the second person: an interview method for the science of consciousness. *Phenom. Cogn. Sci.* 5, 229–269. doi: 10.1007/s11097-006-9022-2

Petitmengin, C., Remillieux, A., Cahour, B., and Carter-Thomas, S. (2013). A gap in Nisbett and Wilson's findings? A first-person access to our cognitive processes. *Conscious. Cogn.* 22, 654–669. doi: 10.1016/j.concog.2013.02.004

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., and Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *J. Appl. Psychol.* 88, 879–903. doi: 10.1037/0021-9010.88.5.879

Reid, K., Flowers, P., and Larkin, M. (2005). Exploring lived experience. *Psychologist* 18, 20–23.

Richards, G. (2002). *Putting Psychology in Its Place: A Critical Historical Overview*. London: Routledge.

Ritchie, S. J., Wiseman, R., and French, C. C. (2012). Replication, replication, replication. *Psychologist* 25, 346–348.

Roediger, H. L. (2012). Psychology's woes and a partial cure: the value of replication. *APS Observer* 25:9.

Rosenberg, S. (1969). "The conditions and consequences of evaluation apprehension," in *Artifact in Behavioral Research*, eds R. Rosenthal and R. L. Rosnow (New York, NY: Academic Press), 279–349.

Schoggen, P. (1989). *Behavior Settings: A Revision and Extension of Roger G. Barker's 'Ecological Psychology'*. Stanford, CA: Stanford University Press.

Schwarz, N. (2010). "Meaning in context: metacognitive experiences," in *The Mind in Context,* 1st Edn. eds B. Mesquita, L. F. Barrett, and E. R. Smith (London: Guilford Press), 105–125.

Shapiro, L. (2010). *Embodied Cognition,* 1st Edn. Abingdon: Routledge.

Speelman, C. P., and McGann, M. (2013). How mean is the mean? *Front. Psychol.* 4:451. doi: 10.3389/fpsyg.2013.00451

Tashakkori, A., and Teddlie, C. (2010). *Sage Handbook of Mixed Methods in Social and Behavioral Research*. London: Sage.

Thompson, E. (2007). *Mind in Life: Biology, Phenomenology and the Sciences of Mind,* 1st Edn. Cambridge, MA: Harvard University Press.

Thompson, E., Lutz, A., and Cosmelli, D. (2005). "Neurophenomenology: an introduction for neurophilosophers," in *Cognition and the Brain: The Philosophy and Neuroscience Movement*, eds A. Brook and K. Akins (Cambridge: Cambridge University Press), 40.

Treisman, A. M. (1960). Contextual cues in selective listening. *Q. J. Exp. Psychol.* 12, 242–248. doi: 10.1523/JNEUROSCI.1820-14.2014

Varela, F. J. (1999). "The specious present: a neurophenomenology of time consciousness," in *Naturalizing Phenomenology: Issues in Contemporary Phenomenology and Cognitive Science*, eds J. Petitot, F. J. Varela, B. Pachoud, and J.-M. Roy (Stanford, CA: Stanford University Press), 266–314.

Varela, F. J., Thompson, E., and Rosch, E. (1991). *The Embodied Mind*. Cambridge, MA: MIT Press.

Varela, F. J., Toro, A., John, E. R., and Schwartz, E. L. (1981). Perceptual framing and cortical alpha rhythm. *Neuropsychologia* 19, 675–686. doi: 10.1016/0028-3932(81)90005-1

Wason, P. C. (1971). Natural and contrived experience in a reasoning problem. *Q. J. Exp. Psychol.* 23, 63–71. doi: 10.1080/00335557143000068

Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychol. Rev.* 20, 158–177. doi: 10.1037/h0074428

Wilson, A. D., and Golonka, S. (2013). Embodied cognition is not what you think it is. *Front. Cogn. Sci.* 4:58. doi: 10.3389/fpsyg.2013.00058

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 McAuliffe and McGann. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

REVIEW
published: 28 October 2015
doi: 10.3389/fpsyg.2015.01615

Target definition for shipwreck hunting

Kim Kirsner *

School of Medicine, University of Notre Dame, Fremantle, WA, Australia

The research described in the present article was implemented to define the locations of two World War II shipwrecks, the German raider *Kormoran*, and the Australian light cruiser HMAS *Sydney*. The paper describes the long and complex trail that led through inefficient oceanographic prediction to ambiguous historical prediction involving a single report and on to precise cognitive prediction based on nine reports from more than 70 survivors, a process that yielded a single target position or "mean" just 2.7 NM (nautical miles) from the wreck of *Kormoran*. Prediction for the position of the wreck of *Sydney* opened with wishful thinking that she had somehow reached the coast more than 100 NM away when cognitive analysis of the survivor's reports actually provided the basis for accurate prediction in a position near to the wreck of *Kormoran*. In the account provided below, the focus on cognitive procedures emerged from, first, a review of a sample of the shipwreck hunts, and, second, growing awareness of the extraordinarily rich database available for this search, and the extent to which it was open to cognitive analysis. This review touches on both the trans-disciplinary and the cognitive or intra-disciplinary issues that so challenged the political entities responsible for supervising of the search for the wrecks of *Kormoran* and *Sydney*. One of the theoretical questions that emerged from these debate concerns the model of expertise advanced by Collins (2013). The decomposability of alleged forms of expertise is revealed as a fundamental problem for research projects that might or might not benefit from trans-disciplinary research. Where expertise can be decomposed for operational purposes, the traditional dividing lines between experts and novices, and fools for that matter, are much harder to discern, and require advanced and scientifically informed review.

Keywords: shipwreck hunting, error, memory, decision making, cognition, mental models, trading zones

OPEN ACCESS

Edited by:
Marek McGann,
Mary Immaculate College, Ireland

Reviewed by:
Martin Lages,
University of Glasgow, UK
Amanda Jane Barnier,
Macquarie University, Australia

***Correspondence:**
Kim Kirsner
pkirsmer@bigpond.net.au

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *18 June 2015*
Accepted: *06 October 2015*
Published: *28 October 2015*

Citation:
Kirsner K (2015) Target definition for shipwreck hunting. Front. Psychol. 6:1615. doi: 10.3389/fpsyg.2015.01615

CONTEXT

HMAS *Sydney* and HSK *Kormoran* sank within an hour or two of each other and approximately 13 nm apart on November 19th, 1941. The British light cruiser and the German raider met by chance while *Sydney* was steaming south from Sunda Strait to Fremantle and *Kormoran* was searching for merchant targets before laying mines off the small coastal port of Carnarvon. The vessels met on a clear afternoon and sighted each other at a distance of 20 or more nautical miles (NM). *Kormoran* turned to the west to avoid combat but *Sydney* followed, and, when *Sydney* closed to less than one NM, combat was inevitable. *Sydney* had squandered her advantages in regard to long range gunnery, director control, armor, and speed. *Kormoran* fired first and the engagement

lasted less than 30 nm. *Sydney* was hit by at least fifty 155 mm rounds, hundreds of smaller missiles, and one torpedo, and she sank with the loss of all hands about 5 h later. Kormoran was hit by only three or four six inch rounds but one of those destroyed her motive power and she was scuttled about 6 h after the battle following an orderly disembarkation of the majority of her crew in five lifeboats and two life-rafts. A brief history of the event was published by Gill (1957, 1985).

PERFORMANCE CRITERIA

Before the detailed analyses are considered, it is appropriate to identify critical performance criteria for the domain, and to underline the relationship between the performance criteria and the author's focus on the mean. The author has adopted three performance criteria, and one simplifying convention.

The first and most important criterion corresponds to the aim of this edition of *Frontiers*, and the focus on the challenge of identifying an optimal search target or *mean* given six or more forms of evidence, the impact of time on the accuracy of each of those forms, and the inevitable presence of human error. The convention involves the use of "Distance from the wreck of *Kormoran*," or *Error*, to minimize reliance on the two-dimensional world of traditional cartography. Distances are specified in nautical miles (NM), where one NM = 1.85 km or 1.15 statute miles. The primary challenge for wreck-hunters involved extraction of a mean target position from the reports available for a particular wreck. The first criterion therefore involved *Accuracy*.

The second criterion involved selection of an *efficient* search box, a box that must therefore include the wreck of *Kormoran* while minimizing the size of the search area. The search areas associated with the historical shipwreck searches of interest ranged from 100 Square Nautical Miles or SNM to 600 SNM, however the areas originally tabled for the search for *Kormoran* involved far larger areas than that, up to 13,000 SNM or more in some cases.

The third criterion involved the extent to which a particular solution reflected the power and variety of the available evidence. All other things being equal, a recommendation that reflected one report and one report only must be set aside in favor of a recommendation that reflected several reports or even a substantial fraction of the available evidence. This approach highlighted the weaknesses associated with cherry-picking. For convenience, this criterion is referred to as *Explanatory Power*.

DISCIPLINES AND EXPERTISE

The question under review in this paper concerns the location of the wrecks of *Kormoran* and *Sydney*. In retrospect, and with the benefit of hindsight, it is now evident that many of the critical entities in the search were overwhelmed by the shear variety and the depth of the evidence available. The critical issue bears some comparison with the signal detection challenge described by Tanner and Swets (1954) more than half a century ago. But there is another problem. Although Thagard (2005) described the boundary regions between disciplines as a critical venue for innovation in science, the absence of informed scientific leadership among the entities responsible for *management* of the search created an unsympathetic environment for science in general and scientific innovation in particular.

Figure 1 identifies seven data types and four or possibly five disciplines with an interest in the search for the wreck of Kormoran. The presence of so many interested disciplines reflected the shear variety and the volume of the known and potential sources of data available for the search. The following is a short summary of the available types of evidence and sources:

1. *Flotsam (Oceanography)*: The first type of evidence involved the positions of flotsam, information open to hindcasting to reconstruct the point or points of entry into the water. However, oceanographic hindcasting depends critically on an understanding of the direction, velocity and stability of wind and water currents, and the increasing challenge faced by hindcasting with the passage of time, where time for this search ranged from 84 to 209 h.
2. *Lifeboat diary (Oceanography and Navigation)*: One person in one of the lifeboats maintained a simple diary recording performance data, evidence that enabled reconstruction of the position of *Kormoran*. In practice, interpretation of the diary depended on *oceanographic* as well as *navigational* assumptions, and, if the former are misunderstood, *navigational* reconstruction can be far off the mark.
3. *Reports from Kormoran survivors (History and the Cognitive Sciences):* It is now apparent that the *Kormoran* survivors provided more than 70 reports about the absolute or relative position of *Kormoran*. In addition, RN and RAN servicemen provided nearly 50 summary reports that included comment about the location of the wreck.
4. *Reports from observers on the coast (History and the Cognitive Sciences):* Commencing with journalist Bryan Clark in the 1980's, more than 90 reports were accumulated from about 30 people living on the coast between Geraldton and Dirk Hartog Island.
5. *Magnetic Anomaly (Geophysics*): The first WAMM/RAN search in 1984 was driven by the presence of an anomaly off the coast near Kalbarri, about 130 nm from the wrecks, and received no support from any other source.
6. *Map Dowsing:* Commencing in 1989 Lindsay Knight and Warren Whittaker claimed that a combination of hand-based and electronic-based map dowsing procedures had located the wrecks of *Kormoran* and *Sydney* near the Abrolhos Islands, 180 nm from the position of the wrecks.
7. *The United States Navy (USN)*: Mike McCarthy, Curator of the West Australian Maritime Museum (WAMM), sought assistance from the USN subsequent to the 1991 Oceanography Workshop. The following quotations are from a FAX from the Curator to David Gallo of the Woods Hole Oceanographic Institute (WHOI) in Falmouth, Massachusetts in 1992:

> "My hopes for the search now lie in anti-submarine warfare records, for it has long been my understanding that many of the magnetic anomalies on the seafloor throughout the world are known and have been mapped for strategic

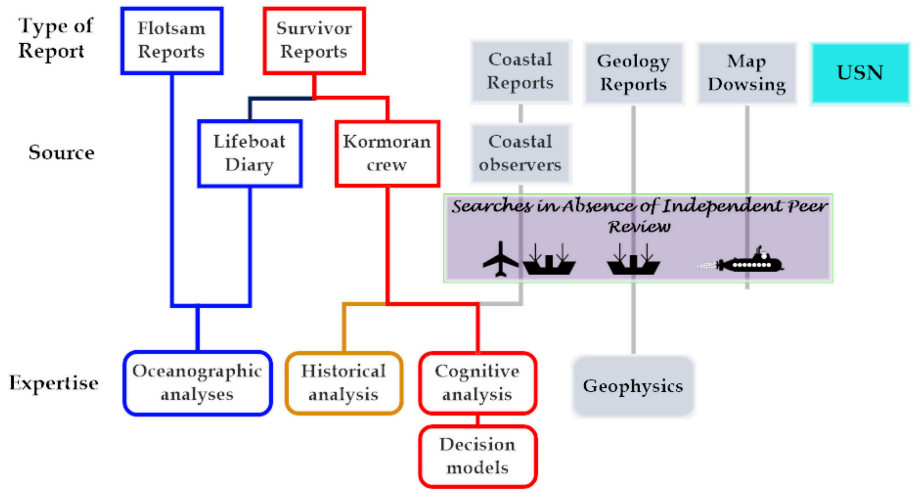


FIGURE 1 | Overview of the disciplines that contributed to the Search Definition debate. The figure indicates the source of each type of report, and the form or forms of expertise relevant to the report types, where known. The blue, brown, and red connections reflect the complexity of the relationships among the reports and the disciplines. The gray boxes identify unproductive streams of argument. The aviation and maritime symbols reflect the focus of the WAMM and/or the services in regard to research and search activities.

purposes. These suspicions have been long since confirmed in discussions with the US, GB, and Australian anti-submarine operatives and were first mooted here in the searches for the SS Koombana many years ago."

And significantly,

> "If the approximate locations of the Sydney/Kormoran are to be found by that route, my problem will be how to keep confidential my source and yet not pretend that we had found the wrecks purely by our own means."

The overview of the disciplines involved in the search acknowledged the extraordinarily rich mixture of evidence, expertise, wishful-thinking, and fantasy that dominated the first 25 years of interest in the search for *Kormoran* and *Sydney* as well as the challenge faced by the private and government entities that engaged in supervision of the search, a challenge they accepted without deploying, seeking, or recognizing the need for expertise.

OCEANOGRAPHIC AND NAVIGATION ANALYSES

In 1991 the author approached the WAMM, and proposed that it design and establish an oceanography workshop, the objective of which was to adjudicate between the positions advanced by Montgomery and Barbara Winter, the trigger for the author's initial interest in the project. The first question therefore involved the power of the oceanographic procedures. Could they be used to adjudicate between the positions advanced by Montgomery and Winter?

The rationale for the position advanced by Winter was clear. Winter (1991) included translations of critical elements from Detmers' *Battle Summary*, and the entry for 1700 h on November 19, 1941 included the following, "Straat Malakka 111E 26S." Winter had tabled the same general position in 1984 on the map shown at page 160. The position was supported in the earlier publication by reference to the statement by Winter that,

> "Calculations, ignoring some minor variables, show that the end of nautical twilight on 19th November 1941, latitude 26°S longitude 111°E, was 1901G; the time quoted by Detmers, give or take a minute."

The critical issue, as recognized by Winter, involved the distinction between the "private" and partially encoded values in the *Battle Summary*, and the "public" positions provided to the RAN interrogators during the Search and Rescue (SAR) and interrogation processes during and following the SAR operation in 1941. The weakness associated with this report involved the probability that the report was intended to be accurate to only the nearest degree, that is 26°S 111°E, as distinct from the nearest minute, that is 26°00′S 111°00′E. Technically, the former involves an area of approximately 3400 Square Nautical Miles (SNM). Justification for the position advanced by Montgomery (1981) was less clear, and relied on selection of one German report, and a dubious claim about the location of the direct route from Sunda Strait to Fremantle.

The original SAR operation conducted by the Royal Australian Air Force (RAAF) and the Royal Australian Navy (RAN) between November 24th and November 29th 1941 yielded eight reports about the locations of flotsam together with 11 reports about the locations of five lifeboats. Two further reports involved the locations of two life-rafts, but these involved chance meetings with passing vessels. The aerial arm of the operation involved approximately nine systematic searches by Hudson aircraft, searches that were dispersed over an area in excess of 30,000 SNM, but searches that were probably too high to detect anything smaller than a lifeboat. In addition, long-range aircraft

examined specific targets out to sea, and smaller aircraft searched along the coast. The maritime arm of the search involved some nine ships, and focused on the area where the flotsam was observed.

Drifting Objects

Oceanographic reconstruction could be based on some or all of the known positions of the objects discovered after the battle. The objects comprised two life-rafts, three lifebelts, one float, one dog kennel, and one raft, and they were discovered between 84 and 209 h after the battle. The critical questions therefore concerned the elapsed time for each object, the drifting and/or sailing characteristics of that object, and the direction, velocity and, critically, the variability, of the currents and winds for the period. In addition, as each object had individual characteristics, the analysis had to be applied to each object as an independent entity.

The professional contributions to the 1991 Oceanographic Workshop used hind-casting based on the movements of some or all of the objects that left *Kormoran* or possibly *Sydney* between 1800 and approximately 2300 h on November 19th, 1941. The objects were discovered approximately 120 NM north of the now known position of the wreck of *Kormoran*. The hindcasting analyses typically relied on velocity and bearing information for four variables; current, wind, wind-driven current, and leeway. The workshop yielded four professional reports. The report implemented by Search and Rescue expert Hughes (1991) actually included the position of the wreck of *Kormoran* but the center of the search area was 33 NM from that wreck, and the overall area was ~ 7850 SNM. A second, by oceanographers Steedman and McCormack (1991), involved an area of ~ 1000 SNM, but it did not quite include the wreck of *Kormoran*. A third, by Penrose and Klaka (1991), did not include a search area but it did specify a 30 NM long contour that passed within ~4 NM of the wreck. The fourth analysis, by CSIRO expert Alan Pearce (1991) asserted that the amount of variability in the current and wind values for the area precluded accurate prediction. The first three reports are reflected in **Figure 2**.

The challenge posed by Pearce was evident in the current rose re-published by him from the from the KNMI (Dutch) Marine Atlas (See **Figure 3A**). What the figure highlights is the extraordinary variability in the bearing and velocity of the currents for the area. The figure should be considered in the context of **Figure 3B** where the presence of huge eddies is

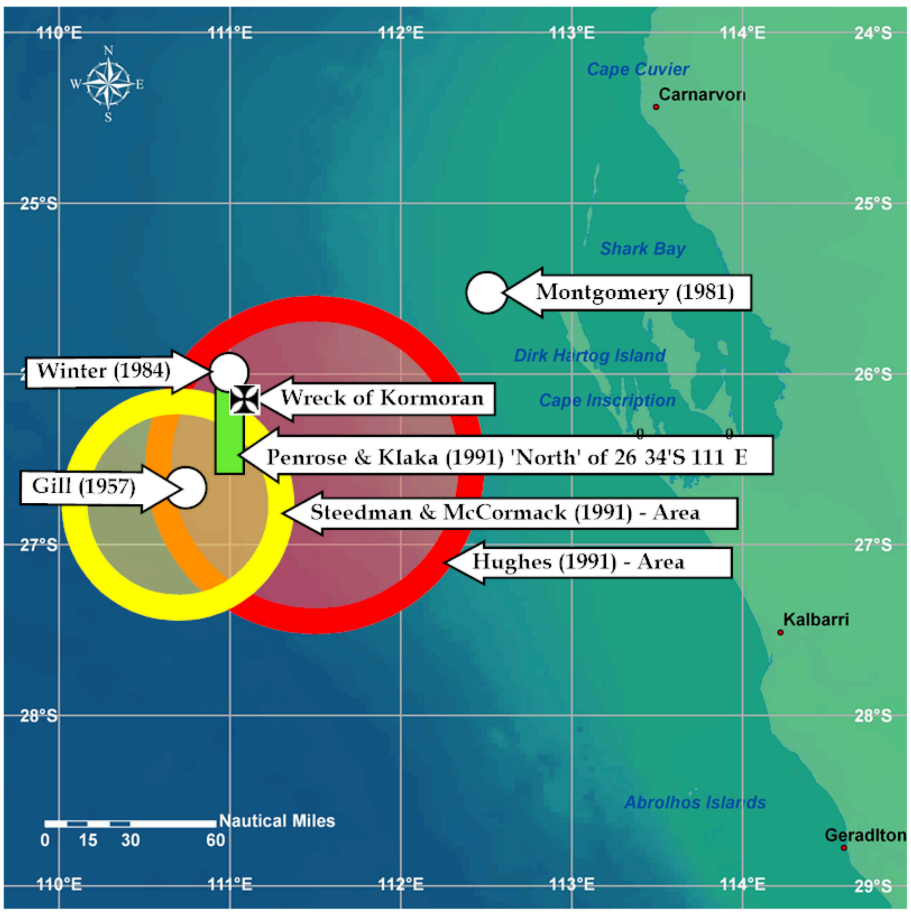
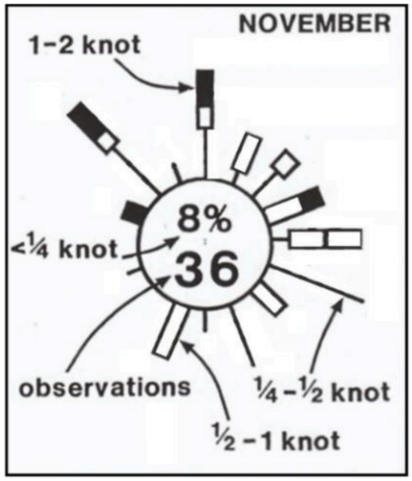


FIGURE 2 | Depicts positions advanced by Montgomery and Winter and results of 1991 Oceanography Workshop. The figure includes the recommendation tabled by Gill (1957, 1985) as well as the now known position of Kormoran.



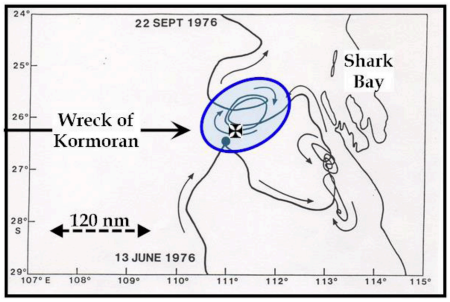


FIGURE 3 | (A) Monthly "current rose" from the KNMI Atlas for area encompassed by 25°S, 28°S, 110°E, and 113°E for November. **(B)** Example of oceanic high showing the scale and the type of movement that complicated prediction off the coast of Western Australia.

noted. The eddies are up to 50 KM in diameter, move in either a clockwise or anti-clockwise direction (for highs and lows, respectively), and the entire system moves gradually from West to East. Furthermore, because the current observations involve almost every point on the compass, the *net drift vector*, the distance made good in any one direction, is very small. Pearce wrapped up his argument in the following terms,

> "It is concluded that "climatological" current data cannot be used with any confidence to predict the likely currents which may have carried debris from the HMAS Sydney away from the site of the engagement."

The role of oceanography for Search Definition was set aside by the author in 1993 for four reasons; first, the size of the error circle defined by Hughes and others was prohibitive for in-water search purposes; second, the argument advanced against reliance on oceanography by CSIRO-expert Alan Pearce underlined doubts about the relevance of the discipline to the search; third, a review that assigned little or no responsibility to oceanography for historically significant searches by Ballard; and, fourth, evolving awareness that the scale and reliability of the reports provided by the survivors might not provide a platform for an efficient search.

The Navigation Argument

The critical note for reconstruction of the lifeboat voyage from disembarkation to the coast is attributed to von Malapert, a member of the crew on the lifeboat captained by Henry Meyer, the navigator. The critical extract from Von Malapert's Diary is as follows:

> (a) ET0-ET71: 12 h drifting, 59 h sailing at an estimated speed of 1.1 knots 062°, Distance sailed 63 miles; (b) ET71-ET90: Drifted in Force 6–7 winds; (c) ET90-ET134; Sailed for 42 h at an estimated speed of 1.9 knot; Estimated distance = 81 miles.

The Navigation argument can be thought of as one facet of the oceanography analysis. Lifeboats, whether under sail or not, are influenced by the direction, velocity, and duration of the prevailing winds and currents. Steedman and McCormack (1991), a professional oceanographer, reviewed, and rejected analysis of the lifeboat journey, arguing that there were too many unknowns about the sailing and drifting characteristics of the lifeboats to accept the diary for formal analysis (Steedman and McCormack, 1991).

In 2000, shortly after the death of Lindsay Knight, owner/developer of the so-called Knight Direct Location System, a map dowser's dream, Warren Whittaker, his long-time collaborator, finally abandoned map dowsing and advanced a new argument for another target near the Abrolhos Islands. The account was based on the diary maintained by von Malapert. According to Whittaker (2000),

> "These "logs" (i.e., written records from the German survivors) contain clear evidence that the battle actually took place west of the Abrolhos Islands and not in the northern or Detmers area. The Abrolhos Islands site is consistent with KDLS Target No. 3 (suspected site of the wreck of HSK Kormoran) at 28°39′S 113°22′E; Error = 196 NM) (Whittaker, 2000).

The claim advanced by Whittaker in 2000 was contradicted by the fact that both Meyer, the Kormoran navigator, and von Malapert, specified the approximate distance covered by their lifeboat over the *entire* voyage, and they put the distance at 150 miles and 153 miles respectively, about half of the distance from the Houtmann Abrolhos Islands to Cape Cuvier by sea.

Whittaker was not the last person to focus on interpretation of the diary. The first endorsement came from LCDR David McDonald RAN. McDonald reviewed and distributed an analysis of the lifeboat voyage that placed the point of origin in a large ellipse off the coast in the latitude of Port Gregory, the latitude long-advocated by his mother on the basis of the oral history accounts described below (McDonald, 2003).

The second endorsement came from the RAN Seapower Centre (Johnstone et al., 2003). The RAN conducted a Lifeboat

workshop in order to facilitate recognition of the site of the wrecks. An expert panel was formed to resolve the issue. The panel's conclusions were as follows:

> "Analysis of the lifeboat voyage by the workshop panel suggests that the correct site of the battle between SYDNEY and KORMORAN lies between the Whittaker and Detmers positions. However, given the paucity of information from the lifeboat log, limited meteorological data from November 1941, and unclear data on the handling characteristics of the lifeboat, the actual position of the battle cannot be narrowed sufficiently to confidently suggest the resting place of the KORMORAN wreck. For these reasons also the Detmers and the Whittaker positions cannot be definitively ruled out at this time."

Consideration of the oceanographic evidence removed the Abrolhos arguments from the table, rendered in-shore locations improbable, and indicated that the discipline could not be used to provide an *accurate* or *efficient* solution.

HISTORICAL ANALYSES

In the first professional historical analysis of the engagement, Herman Gill located the contact and battle positions for *Kormoran* and *Sydney* near 26°34′S 111°E and 26°40′S 110°33′E, respectively, the second of these positions being 42 NM from the wreck. Gill was not of course touched by any interest in a search for the wrecks. As discussed above, Barbara Winter interpreted Detmers' Diary correctly, and located the battle and therefore the wrecks in the vicinity of 26°S 111°E (Winter, 1984, 1991).

The next historical analysis was published by Wes Olson in 2000. Based on the map on page 192 (Olson, 2000), Olson located *Kormoran* near 26°34′S 111°E when she first encountered *Sydney* on November 19th, and the extrapolation developed by Olson placed the battle and therefore the wrecks near ∼ 26°42′S 110°35′E, 42 NM from the wreck of *Kormoran*.

The final historical analysis, by Olson et al. (2001), reverted to the argument advocated by Winter (1984); Winter (1991), however their paper included a new and independent detail. They assumed that 26°34′S 111°E specified *Kormoran's* noon position, and used dead reckoning based on Detmers' account of *Kormoran's* subsequent movements to locate the wreck near 25°58′S 110°56′E. This position is 11 NM from the wreck of Kormoran, however, as they advocated a search circle with a radius of 10 NM only, their analysis did not quite include the wreck of *Kormoran*.

The historical analyses focused almost exclusively on the content and interpretation of the report included in Detmers Battle Summary. Before turning to the cognitive analyses, brief consideration will be given to the technologies that so captured the attention of the WAMM, the RAN, and the public agencies engaged in the search between 1981 and 2005.

The original historical argument provided a target accurate to only 30′ or 27–30 NM, thereby defining an area of ∼ 3400 SNM. When combined with dead reckoning a more specific target was provided but in each case the solution relied on only one report and one source, a source that had provided inaccurate and inconsistent information at the time of the in-water search after the engagement.

MAGNETIC ANOMALIES, MAP DOWSING, AND ORAL HISTORY

The amount of credence placed on Montgomery's claim by the WAMM is evident in the fact that the relevant team used it to justify an in-water search involving collaboration with the RAN while searching to the south of 27° South, more than 130 nm to the south of the position actually recommended by Montgomery, and the position recommended by Winter, and the wreck of Kormoran (See Green et al., 1984).

The Map Dowsing argument passed through two incarnations. The first of these involved traditional hand-based map dowsing or "divining" while the second was based on the principle of Electron Spin Resonance. Each of these procedures pointed to positions off the Abrolhos Islands, nearly 200 NM from the now known positions of the wrecks. One of the positions allegedly attracted a search from an RAN submarine in 2000. In 2003, based on Whittaker's interpretation of von Malapert's diary of the voyage of a lifeboat, the RAN established a workshop involving four senior navigators, and provided qualified endorsement for the navigation argument for the Abrolhos Islands.

The Oral History argument reflected an interesting and recently established branch of history, however it is usually used to capture *subjective* experience as distinct from fine details about the timing or the location of specific events. For example, Studs Terkel, a key player in the tortuous history of the discipline, noted that,

> "They would sit around and tell us their hard luck story. Whether it was true or not, we never questioned it. It's very important you learn people as they are. At that particular moment when you are talkin' to that person, maybe that's how that person were. Tomorrow they can be different people." (Emma Tiller, *a cook in Western Texas, as reported by Studs Terkel, 1970).*

In fact, the majority of the oral history reports submitted as evidence of a battle near the Abrolhos Islands involved eye-witness accounts by individuals, and the claim that they involved Oral History was therefore problematic. As accounts based on Remote Memory, they involved critical flaws. For example, Bryan Clark, the journalist who first recorded many of the stories in the late 1980's, opined that some of them at least reflected experiences from later years, when Catalina maritime patrol aircraft flew practice missions off Port Gregory. In our view (Kirsner and Dunn, 1998c), the accuracy levels for recall of remote events are so low after an interval of nearly 50 years that little confidence can be placed in them (e.g., Wagenaar, 1986). Furthermore, the historian providing advice to the *Joint Standing Committee on Foreign Affairs, Defence, and Trade* (JSCFADAT) advised that few of the accounts included any form of link to the engagement between *Kormoran* and *Sydney*. Statistical analysis supported this argument and revealed that fewer than 10% of the accounts actually included any link with *Kormoran* and *Sydney* at all. Another line of argument indicated that the reports emerged from positions covering more than 20,000 SNM, hardly a pointer to a specific battle on a given day. None of these arguments prevailed. The JSCFADAT gave the oral history argument pride

of place for the 2001 Shipwreck Seminar; and the RAN and the RAAF implemented expensive and risky aerial and surface searches of the target positions identified by the oral historian.

Analysis implemented by the author and his colleagues rejected the remote memory or "oral history" argument, despite strong support from WAMM and the JSCFADAT (Kirsner and Dunn, 1998c), and the map dowsing case was intrinsically weak (Kirsner and Dunn, 1998b).

THE COGNITIVE SCIENCES

Review of Wreck-Hunting

Following the Oceanography Workshop, and aware that oceanography would not be able to provide a precise target, the author sought to achieve a better understanding of the challenges and solutions associated with deep-sea wreck-hunting. The first section involved a review of the available evidence on deep-sea/off-shore shipwreck hunting, with the focus on the identification of search targets and the definition of search areas. The critical history of the engagement between *Kormoran* and *Sydney* was published by Hermon Gill in the first book of his two volume history of the RAN in World War II (Gill, 1957, 1985). Gill wrote more than 12 pages about the engagement between *Sydney* and *Kormoran,* however, unavoidably, his analysis was based exclusively on reports provided by the *Kormoran* crew, and that was perhaps the first trigger for doubt among the old salts in the local community. Further doubt about the reliability of the reports provided by the *Kormoran* survivors was facilitated by the fact that the Captain and the Navigator provided inconsistent reports to the RAN officers during late November and early December 1941. A second cause for doubt arguably involved the gradual release of information about the role of Signal Intelligence following World War II, a process that was still yielding new information and an occasional surprise up to the very end of the twentieth century (e.g., Fry, 2012). A third issue that compromised the search debate between 1991 and 2013 involved the widespread assumption by the official bodies associated with the search that a sou'wester was the key to expertise, and that scientists without sou'westers had no business entering the arena.

The first section of the review involved consideration of five examples of deep-sea wreck-hunting. The first example involved the search for the wreck of the *Titanic*, sunk on April 14th–15th 1912. The search area adopted for the first three searches for *Titanic* appear to have been constrained exclusively by navigational reports about the position of the sinking, and the resulting search involved only 100 SNM. In 1985 the area was expanded by Ballard to 150 SNM to incorporate the southerly movement of the lifeboats between the sinking and the rescue but even here the critical factor involved navigational reports about the final position of the lifeboats (i.e., without reference to oceanographic assumptions), coupled with a decision to commence the search beyond even that position, and shape the in-water search from that position toward the estimated position of the wreck of *Titanic*. In summary, the critical points were determined solely by navigational reports although the reports were selected to define a search area that reflected the movement of lifeboats in the water. The current in-water technology enables more efficient in-water search, but that should not be critical if the actual search box has been chosen with due consideration for uncertainty.

The second example involved the German battleship *Bismarck*, sunk on May 26th 1941. The search area for *Bismarck* was shaped around reports about the sinking position provided by British battleships HMS *King George V* and HMS *Rodney*, and British cruiser HMS *Dorsetshire*, although only the third of these was present when *Bismarck* actually sank. As the search unfolded however the focus shifted to a search for debris, and then a landslide on an underwater mountain, the end of which finally revealed the location of the wreck. The assumption adopted by Ballard was that the landslide was actually triggered by *Bismarck*, when it hit the ocean floor. Ballard indicated that the search area involved was = 200 SNM. Descriptions of the search operations for these wrecks are detailed in Ballard (2008), and available from earlier reports by Ballard (1988, 1990); Ballard and Archbold (1999).

The third example involved the US Aircraft Carrier *Yorktown*, lost during the Battle of Midway. A review of the search indicated that a search area of ≤500 SNM was used by Ballard, and that the area was specifically extended to the south in order to cater for uncertainty about the distance covered by Yorktown between the final aerial attack on the afternoon of June 5th and her sinking on the morning of June 7th following a submarine attack on June 6th.

The fourth example involved the search by David Mearns for the bulk carrier *Derbyshire*. Initial analysis revealed three reliable reports of oil slicks. Further, analysis suggested that the wreck might be up to five nm to the north of the position where the oil actually breached the surface. Mearns (1995) defined two search areas, of ∼ 90 and ∼ 170 SNM as "high" and "low" probability areas respectively, and the wreck was duly found in the predicted area. It is a matter of interest that Mearns used "the principles of modern probability analysis" as described by Discenza and Greer (1994) to shape the search plan.

The fifth example involved the search by Mearns for the wreck of HMS *Hood*, sunk on May 24th, 1941. Information about this search was not available in the public domain until 2001, and the work did not therefore inform the author's review. As described by Mearns and White (2001) however, the record included no fewer than 10 reports about the location of *Hood*. Mearns rejected three of these because they depended on aerial calculation. Of the remaining seven no fewer than five were from battleships or cruisers and occupied a very tight box of approximately 40 SNM. The remaining two involved positions determined by destroyers and either dead reckoning or movement after a substantial time lag (and therefore uncertainty over wind and current). Mearns tabled two search boxes for operational purposes, of ∼ 600 and 200 SNM respectively, however the quantitative bases for these areas remain unspecified. Mearns and White (2001, p. 107) noted however that,

> "The first two decisions were dictated by the simple application of the navigational errors we had found to exist in the reported sinking positions of the reported sinking positions during the

> First and Second World Wars. The errors that I chose to apply in this case were divided into two different categories: the worst error reported by a surface ship and the average error reported by a number of surface ships. These circles of error were drawn around each of the five most likely sites for Hood to have sunk."

The details of the in-water searches conducted by Mearns have not been published, as they formed "part of a commercial operation."

The history-based procedures implemented by Ballard and Mearns realized substantially smaller search areas than those generated by the 1991 Oceanography Workshop. The three searches by Ballard involved areas that ranged 150–500 SNM, values dramatically smaller than those generated by the 1991 Oceanography Workshop, and areas that enabled discovery of each of the wrecks concerned. The areas used by Mearns in searches for the SS *Derbyshire* and HMS *Hood* are less clear but they were probably less than 600 SNM, and they too relied primarily on contemporary reports from observers.

The review removed any doubt about the relative merits of the oceanography-based and history-based procedures in research to define accurate and efficient areas for in-water search. The oceanography-based procedures yielded an overall area of ∼8400 SNM for *Kormoran* [sum of areas provided by Hughes (1991) and Steedman and McCormack (1991)], although even that solution came with a significant caveat from CSIRO based expert Alan Pearce. The central issue was therefore clear. As the oceanography-based analyses for *Kormoran* had produced search areas between 10 and 100 times larger than the areas used for the *Titanic*, *Bismarck*, and *Yorktown* searches, an historically-based analysis was essential, and the author embarked on the collation and analysis of the survivors' reports.

The review indicated that search definition was dominated by reports from captains, navigating officers and professional observers, and that it generally resulted in areas of 500 SNM or less.

The Kormoran Database

The critical question concerned the scope, extent, and reliability of the reports provided by the German survivors. Given inconsistent reports from the Captain and the Navigator, was it possible to accept as valid reports from other crew members, particularly if they too varied from report to report? The records at WAMM provided an initial set to work on, and the books published by Montgomery (1981) and Winter (1984) provided pointers to additional material, however it was by no means obvious that these sources covered the full extent of the reports provided by the Kormoran survivors and the RN/RAN interviews and interrogations.

The second step involved archival research in London, Washington, and Norfolk, Virginia as well as Sydney, Perth, Canberra, and Melbourne, in Australia. When combined with the material available from Fremantle, the archival research yielded a total of 73 reports that involved reference to the absolute or relative location of *Kormoran*, a further nine about the bearing and distance of *Sydney* relative to *Kormoran* for the period between the battle and the last sighting of *Sydney*, and a further 44 that involved official or unofficial reports from RN or RAN officers. Collation of the reports and the creation and analysis of a substantial database located the project firmly within the tradition of error analysis in Cognitive Psychology and Human Factors. The project therefore required consideration of two data types, involving the positions of objects in the ocean and the reports of the survivors, and three methodological approaches, involving oceanographic hindcasting, historical review, and cognitive analysis and modeling.

The products of the archival research were summarized in the following extract (Kirsner, 1997b). The paper was entitled *The War of the Ghosts: Using dusty records to hind-cast the locations of HMAS Sydney and HSK Kormoran* and it was presented to a Humanities Conference at the University of Western Australia.

> The traditional problem with archives is that they contain too little information, and that too many inferences must therefore be left to logical analysis or intuition. The archives concerning the loss of *Sydney* and *Kormoran* arguably involved the opposite problem where location is concerned. Analysis of the archives and other historical sources revealed at least 60 separate sources of information about the location or locations of the wrecks, and these sources identified no fewer than 25 different sites, only a few of which could be discounted absolutely. The sources are, furthermore, distributed among five or six layers involving SAR operations, the interrogation of survivors both during and after the SAR operation, operational reports prepared by RN and RAN officers, administrative reports, political reports and, finally, historical argument. Worse still, the deeper layers even include reports suggesting new sites, not recorded in the earlier reports.

The data depicted in **Table 1** is a summary of the reports from the Kormoran Database. The reports were obtained from numerous sources. Some of these were available from Montgomery (1981) and Winter (1984); some were from the library of the West Australian Maritime Museum; some were obtained from the state archives in Perth, Melbourne, Sydney, and Canberra; and a handful were discovered in the national archives of the UK and the USA, and two were discovered by Hore and Mearns in the Old Admiralty Library in London. Most of the reports were collated between 1993 and 1997, however additional items were added later as they became available. A summary file was provided to the Cole Commission at its request in 2008, and re-distributed by it on request.

Table 1 can be read as a form of "stem and leaf" diagram. The numbered reports in the fawn rows were included in the final analyses; the reports with black bullets were treated as derivatives, and discarded; and the bullets with open circles were treated as outright errors.

The Kormoran Database comprised more than 70 reports by survivors about the location of the wreck of Kormoran, a source of evidence that would be invaluable for an accurate and efficient solution provided that the major part of the database was reliable. A substantial database was essential if the solution was to be efficient as well as accurate.

Reliability of the Kormoran Database

Figure 4 is a plot of the data from **Table 1**. The axes depict the data in Log-Log coordinates. The y-axis reflects a log transformation of the number of reports associated with each Type of Report. The x-axis reflects a log transformation of the

TABLE 1 | Stem and leaf plot of Reports from Kormoran survivors.

Type of Report	AR	E	Comment
1. 26°S 111°E (to be read as ±30′)	17		Mode (including one report from Bunjes, two from Detmers)
• 26°S 110°E		2	EC including one report from Detmers
• 26°S		1	EO
• 26°S 11°E		1	EO
• 26°S 108°E		1	EC
• 26°S 111°40′E		1	EC
• 26°S 111°21′E		1	EC
• 24°S 111°E		1	EC
• 25°S 111°E		2	EC including one report from Detmers
1a. 26°S 111°E (to be read as ±30′)	1		Meyer as revealed by diary in 2000
• 27°S 111°E		5	Initial reports from Meyer
• 26°30′S 111°40′E		1	Later report from Meyer
2. 120 nm from Coast	4		Bunjes; 120 nm selected on basis of MDP
• 150 nm from coast		2	EC
• 60 nm off land		1	EC
3. 160 nm SW of Cape Cuvier	0		Bunjes; Cape Cuvier selected on basis of MDP
○ 160 nm SW of *NW Cape*		1	
4. Geraldton signal 2 (gap) 7S 11115E	1		26°07′S 111°15′E selected on basis of MDP
○ 7C 115E 1000 GMT		1	
5. Sailed 150 nm NE to land	1		Meyer—lifeboat diary
• Sailed 153 nm NE to land		1	V Malapert—lifeboat diary
6. 26°34′S 111°E 3	6		Detmers: Winter (1991) classified as *noon* report
• 26°32′S 111°E		6	Detmers
• 25°34′S 111°E		2	Detmers
• 26°31′S 111°E		1	Detmers
7. 130 nm SW of Shark Bay	4		Habben
8. Due West of Shark Bay at 2000 h	1		Detmers to be "due west of Shark Bay at 2000"
○ 120 nm SW of Fremantle		4	EC
○ 100 nm off Fremantle		1	EC
○ 130 nm due West of Perth		1	EC
○ 125 degrees SW of Frem.		1	EC
○ 20 nm SW of Fremantle		1	EC
Total	35	38	

The numbered and colored items define the stems. The bulleted items comprise the leaves associated with each stem. The gray items are "pure" errors.

rank value of each Type of Report. Following Zipf (1949), the resulting function is linear and the observed pattern is consistent with the proposition that resource limitations played a role in report selection and recall. Zipf demonstrated that a linear relationship is observed for many relationships provided that Log-Log axes are used. The r^2-value for the survivors' reports was 0.89, a value that accounts for more than 80% of the observed variation in the data. A small sample of the many variables that honor Zipf's Law includes word frequency distributions for English (Zipf, 1949; Miller and Newman, 1958), recall (Kaplan and Carvellas, 1969), and character frequency distributions for Japanese Kanji (Speelman and Kirsner, 2005). The figure also depicts the equivalent set of results for the 44 reports tabled by RN and RAN officers, and it shows essentially the same pattern.

Zipf's Law does not constitute a "proof" in the definitive sense of that term. Rather, it is a pattern we would expect to observe in a memory study involving recall of a number of words. The distributions of the reports are consistent with the proposition that the observed patterns reflected randomly distributed memory or transcription errors. The most obvious alternative hypothesis involved the argument that the Kormoran survivors rehearsed their answers. In the extreme case, this approach would have produced just one Type of Report, or something close to that. The fact that the reports were distributed across eight or more than eight referents contributed to the assumption of reliability.

Two issues were critical to the final assessment; first, the fact that no fewer than eight independent groups of reports or *constraints* pointed toward the same general position, and, second, the fact that all five lifeboats either arrived at or were approaching one point on the coast when they were discovered by rescue craft, a degree of convergence that would have been

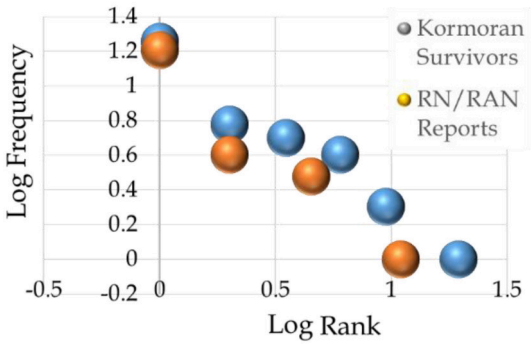


FIGURE 4 | Zipf functions for the Kormoran survivors (blue circles) and RN/RAN officers (red circles) associated with the 1941 interrogations and interviews. The figure depicts Log Frequency plotted against Log Rank for each set of data.

improbable had the survivors had no idea where they were, or where they were going.

Mathematical analyses determined that the Database was reliable, overall, although it was clear that a number of individual reports were not.

The 2004 Solution

The majority of the reports used for the wreck-hunts described in the review relied on reports from navigators about the coordinates of the vessel prior to or at the moment of loss. Each set of reports involved some variability among the navigators who reported the loss of a vessel, involving the own ship navigation, time of the observation, or the position of the survivors, however they all relied on professional reports, where precision was reduced because each navigator produced a unique solution. The *Kormoran Database* reflected a very different form of evidence and error. Many of the reports provided only a *general* guide to the location of *Kormoran* at the time of her loss, for sailors at sea, and at risk, and an alternative approach therefore involved a weaker assumption that all or at least many of the reports were valid, and could therefore be considered as a set to point to the position of the wreck. Working alone in the first instance, and then in collaboration with John Dunn, this approach was refined in four stages. The stages were as follows:

1. Discount and remove obvious errors from consideration.
2. Group reports that involved a single concept, or "root," in evolutionary terms.
3. Develop principles to resolve competition when reports in a single group involved inconsistent evidence.
4. Design and implement a mathematical decision model to integrate the surviving statements or constraints, the task completed by John Dunn.

The Constraints

The overall procedure was designed to produce a single and accurate estimate of the location of the wreck of *Kormoran*. The analysis evolved over the period 1991–2004. The concept of *converging operations* shaped the research.

Constraint 1: 26°S 111°E

The majority of the 18 reports that involved 26°S 111°E were provided by Wireless Telegraphy Officers (WTOs), adding further weight to the validity of the report. The critical weakness with the mode is that the position as reported, 26°S 111°E, is accurate to only the nearest degree, and for wreck-hunting purposes it should therefore have read 26° ± 30′S 111° ± 30′E, where provision for error identified a search area of 3400 SNM.

Constraint 2: Report by Bunjes that the battle occurred "120 nm from the coast"

The second constraint involved the distance from the coast. Three estimates were available from the reports from the Kormoran survivors, at 60, 120, and 150 nm. One hundred twenty nanometer was adopted for two reasons: First, Bunjes provided 120 nm value on three occasions during the fortnight after the battle whereas he provided the value of 150 nm in one report only, and years after the event; and, second, 120 nm provided a better fit with the first and third constraints under the Minimum Distance Principle described below.

Constraint 3: 160 nm SW of NW Cape (interpreted as Cape Cuvier)

The third constraint, also attributed to Bunjes, involved the report that the battle occurred 160 nm "South-West of North-West Cape." North-West Cape is more than 300 nm from the area of the battle, and out of the game. An error is the obvious explanation but what sort of error. While the author was working out the tracks of the lifeboats in 1991 (Kirsner, 1991), detailed analysis indicated that all five of the lifeboats could have been heading for the same position on the coast. Two of the lifeboats—those captained by Meyer and Kohn–beached 5 and 12 nm north of Cape Cuvier respectively, and the other three were sailing east and more and less directly for Cape Cuvier after some 4 or 5 days drifting to the north with the current and wind.

Triangulation

Given the availability of multiple and converging constraints, triangulation provided an appropriate model for our approach to the problem. Maritime triangulation is illustrated in **Figure 5A**. In that example the approximate location of a ship is assumed to be inside the triangle defined by convergence among the three observations or "Lines of Position" specified by the navigator. **Figure 5B** is a summary of three reports provided by Wilhelm Bunjes, a sometime officer in the pre-war German merchant marine. Argument for the reliability of Bunjes' intentions came from the fact that one of his reports about the Kormoran officer's was "masked" in the archives for nearly 30 years, allegedly to protect him from repercussions associated with his anti-Nazi sentiments.

Although the Kormoran Database included reports that relied on a variety of referents, including cartographic coordinates, distance from the coast, and distance and bearing from coastal features, it is evident that Bunjes' reports converged on a

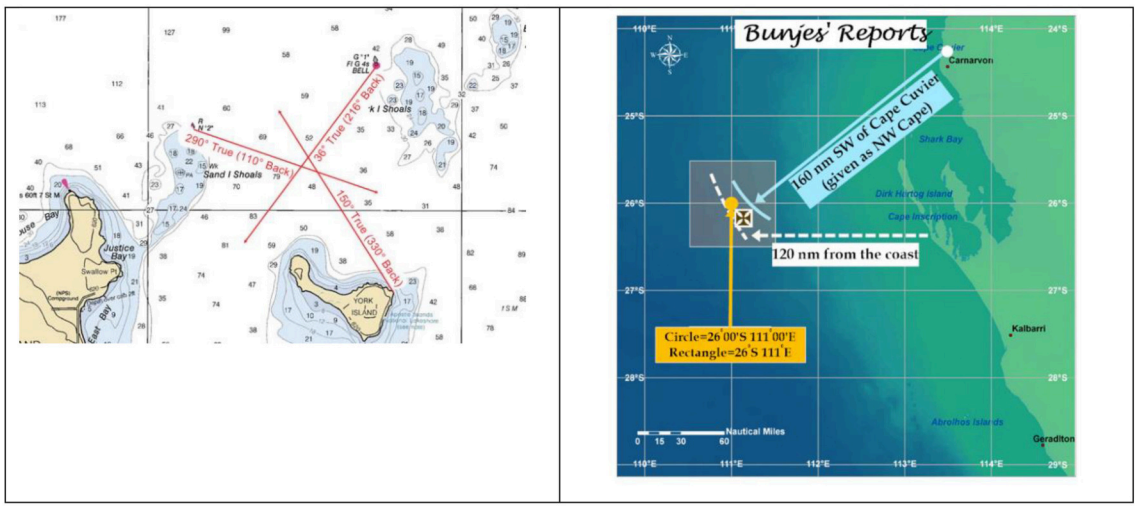


FIGURE 5 | (A) Illustration of triangulation as the term is generally understood by mariners (Hansel, 2011). **(B)** Convergence among reports from Bunjes. The circle and the rectangle represent 26°00′S 111°00′E, and 26°S 111°E, respectively.

single "position," a position that could be used to define the point of disembarkation from Kormoran prior to her sinking. Convergence was not straightforward. As indicated above in regard to the first three constraints, convergence was achieved only when Cape Cuvier was substituted for North-West Cape as the coastal feature, and 120 NM was adopted in preference to 150 and 60 NM as the distance from the coast, and even then the "error" associated with the first constraint a significant handicap.

The authors' use of triangulation is closer to its nautical roots than it is to the role of triangulation in the social sciences (e.g., Yeasmin, 2012). In the latter case it constitutes a form of validation although it may also be implemented in order to increase "understanding" of a specific problem. In the present case however, triangulation is being used to refine the location of a wreck by using "Lines of Position" when most if not all of the lines involve an element of potential but unknown error and uncertainty.

The approach outlined above involves very different principles and assumptions from the oceanographic models, however it is the contrast with the historical analyses that is particularly interesting. Six historians or teams of historians tabled solutions to the Search Definition problem; Gill (1957, 1985); (Error = 42 NM), Winter (1984); Winter (Error = 7 NM), Winter (1991); Winter (Error = 7 NM), Olson (2000); Olson (Error = 42 nm), Olson et al. (2001, Error = 11 nm), and Hore and Mearns (2003); Hore and Mearns (Error = 7 NM), and in each case attention was restricted to a single option or interpretation of one or possibly two reports. It is evident that the attention of the historians was focused either substantially or exclusively on the reports provided by Detmers, the Captain of *Kormoran*, and that the only issue that vexed them concerned the relative merits of the *noon* and *battle* interpretations of 26°34′S 111°E. Indeed the only individuals or teams to opt for that remote position as the position of the *battle* and therefore the wrecks were Gill (1957, 1985), Olson (2000), and Mearns (e.g., Finding Sydney Foundation, 2007), and Mearns included the so-called *noon* position, 26°34′S 111°E, in the in-water search area in 2008, a decision that depended on his recommendation alone.

Constraint 4: 2#°#7′S 111°15′E; Geraldton signal received at 1800G (interpreted as 26°07′S 111°15′E)

The Geraldton signal has come down to us in two forms. The first form was included in a report prepared by SWACH and dated November 27th. The wording of the report is as follows:

> "Geraldton radio reports that at 1005Z/19/11 they received a weak message. The beginning was unintelligible. Then followed "7C 115E 1000 GMT." The radio operator could not estimate the distance. No Qs were distinguished. They waited 2 min but there was no repetition"

The second version of the report is included in the Fremantle Report of Operations for the period November 24–29th (see Olson et al., 2001, p. 38). The wording of this report is as follows:

> "At about the same time Geraldton radio picked up a weak signal unintelligible excerpt for '2 (gap) 7 111 15 East 1000 GMT (These two reports were not received until 1345H/27)"

The number of operational and cognitive steps between the Kormoran transmission and the SWACH report of operations is difficult to estimate. Radio signals occurred in noisy environments, and it is no accident that signal detection theory (Tanner and Swets, 1954) evolved as a response to the classification problems experienced by radio operators during and after World War II (e.g., Shannon, 1949). We can safely assume that the radio operator in Geraldton was dealing with a noisy signal. She or he may have misheard parts of the signal. They may have heard it correctly but made a transcription error. They may have transcribed the signal correctly, only to have a

supervisor introduce an error, in reading or during preparation of a signal for transmission to SWACH. We do not know for example why the second report comprised "2 (gap) 7" whereas the first comprised "7C," and we probably never will.

The Minimum Distance Principle

The solution adopted to solve the uncertainty associated with this potential constraint involved the Minimum Distance Principle. In brief, six alternative interpretations of the signal were benchmarked against the established candidates; that is, constraints 1, 2, and 3, and the alternative that involved the smallest movement was adopted. The positions in the mix were; 25°37′S 111°15′E, 25°47′S 111°15′E, 25°57′S 111°15′E, 26°07′S 111°15′E, 26°17′S 111°15′E, and 26°27′S 111°15′E. As illustrated in **Figure 6**, the fourth of these positions provided the best fit, and 26°07′S 111°15′E was therefore adopted as the fourth constraint.

Constraint 5: Meyer's "lifeboat originated 150–153 nm SW of landing point"

The critical information is summarized in Section Oceanographic and Navigation Analyses, The Navigation Argument. The uncertainty associated with the relevant estimate was acknowledged by Meyer.

Constraint 6: Report from Detmers' Battle Summary: 26°34′S 111°East

Barbara Winter provided the critical interpretation of Detmers' Battle Summary (Winter, 1991). Her analysis left no doubt that 26°34′S 111°E and 26°S111°E were the *noon* and *battle* positions of *Kormoran*, respectively.

We nevertheless used the distance between the solution offered by this potential constraint and the position provided by other constraints to test the *noon* and *battle* interpretations of Detmers' report. The *noon* interpretation won that competition too, and we therefore adopted the *noon* interpretation for integration purposes. Our dead reckoning analysis confirmed the argument advanced by Olson et al., 2001, and yielded a solution one NM to the North of theirs, 12 NM from the wreck of Kormoran.

Constraint 7: Report by Habben: "130 nautical miles south-west of Shark Bay"

Siebelt Habben, a medical doctor, was repatriated to Germany in 1943 as part of a Prisoner-of-War exchange. Habben provided descriptions of the action between Kormoran and Sydney to the German naval authorities, and the Kreigsmarine subsequently included them in Operationen and Taktiks, Volume 10.

Constraint 8: Detmers' statement that Kormoran should be due west of Shark Bay at 2000G

According to Detmers (1959),

> "The KORMORAN was proceeding at medium speed on her usual sweep and gradually approaching Shark's Bay from the south west. At 1500 h I checked the ship's course and decided to carry on without change until 2000 h, and then turn eastward toward Shark's Bay."

This solution to this constraint also involved dead reckoning, from the assumed track of Kormoran from noon to 1700 h.

Constraint 9: Mathematical analysis identified a "circle of equal speed" for the life-rafts discovered by Aquitania and Trocas

This analysis was submitted to and published by the 2001 Shipwreck Workshop (Dunn and Kirsner, 2001). The mathematical model designed by Dunn was based on three assumptions about the life-rafts;

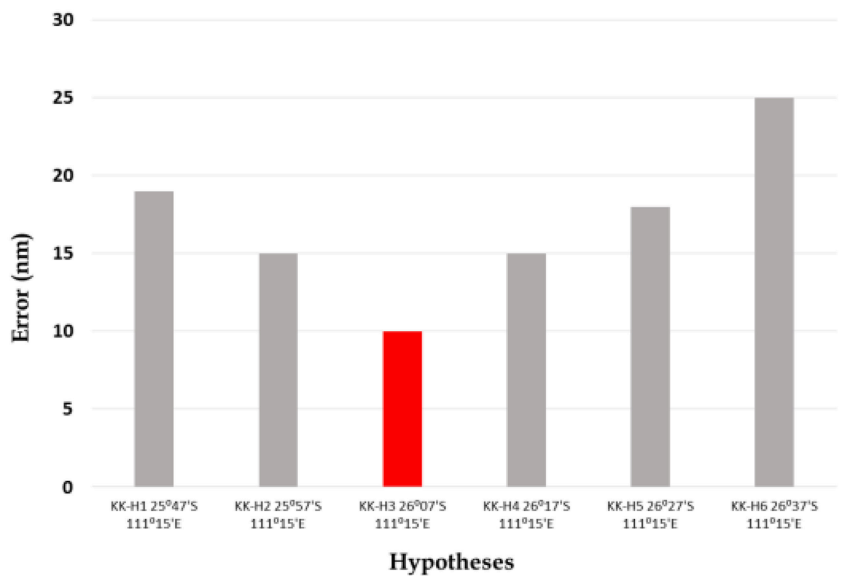


FIGURE 6 | The Minimum Distance Principle. Distance between six candidate interpretations of the signal and the position defined by the first three constraints.

- They were under the influence of the same currents and winds, and they would therefore display essentially the same "sailing" characteristics.
- They had similar buoyancy and "sailing" characteristics, and they would therefore move downwind in a similar direction and at a similar velocity.
- They would conform to wind direction ±35°, an assumption accepted by the Search and Rescue profession.

The cross in **Figure 7** denotes the now known position of *Kormoran*. The distance between Kormoran and the nearest point on the circle is ∼2 nm. The blue circles denote the areas advanced at the 1991 Oceanography Workshop by Hughes (1991) and Steedman and McCormack (1991). The Circle of Equal Distance reflected a purely mathematical solution based only on the assumption that the life-rafts were influenced by the same forces, and drifted at the same velocity. The area of the red circle is irrelevant; prediction involved the circumference.

Integration

In 2004 John Dunn designed a mathematical generalization of the Minimal Distance Principle. The aim of the generalization was to identify the most likely position of the wreck, and the procedure involved selection of the position that involved the smallest "movement" for the set of nine constraints outlined above. We therefore integrated all of the available information under the assumption that each piece of information would be broadly consistent with the remainder, and that integration would converge on the most likely point.

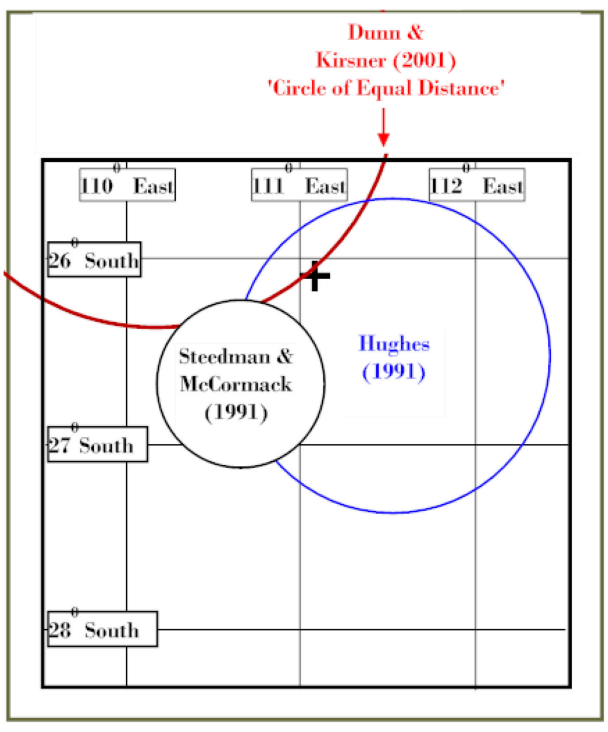


FIGURE 7 | Circle of Equal Distance (Figure from Dunn and Kirsner, 2001).

For each candidate location, corresponding to a point in the ocean, and each constraint, we calculated the minimum distance that the candidate location would have to be moved in order to satisfy the constraint exactly. We referred to this measure as the *error distance* for each location-constraint pair. We then calculated the *average error* distance across the set of constraints for each location which then provided a single *goodness of fit measure* for that location. Clearly, a candidate location with a relatively small average *error distance* satisfies the constraints to a greater extent than a point with a relatively large average *error distance*. No single candidate location satisfied all of the constraints exactly.

Averaging the error distances treats each constraint as having the same weight or importance. We considered and rejected a range of weighting schemes, however we were not persuaded that there was any basis for treating one constraint as more critical than another. We were also guided by studies of expert decision making in which equally weighted linear models (so-called improper linear models) are nearly as efficient as optimally weighted models (Dawes, 1979).

Integration yielded 26°04′S 111°02′E as the position of the wreck. This position is 2.7 nm from the true position of the wreck as established by the *FSF* in 2008 (Finding Sydney Foundation, 2008). The approach was described by Kirsner and Dunn (2004) and Dunn and Kirsner (2011). The recommendation was also used and published by the FSF in 2005 and 2007.

Performance

Accuracy

FSF Director Bob King chaired the Technical Search Committee of the FSF from 2005 to 2007 inclusive. In 2005 King designed a Powerpoint presentation for use by the FSF. The critical figure is reproduced as **Figure 8** below. The figure includes the positions recommended for *Kormoran* and *Sydney* by the FSF in 2005 on the basis of the arguments and recommendations advanced by Kirsner and Dunn (2004). They are depicted as black stars (from the original) identified by the black labels (added) indicating the names of the two ships. The now known positions of the two wrecks are depicted by solid red circles identified by red labels,

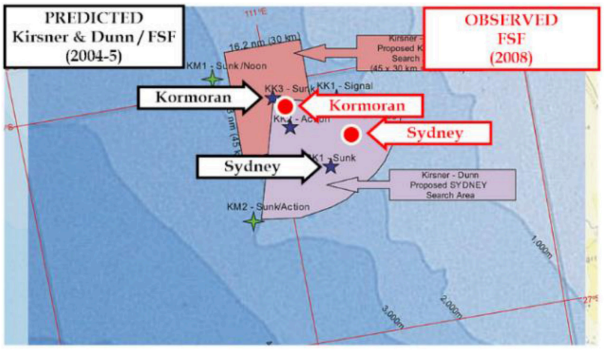


FIGURE 8 | Map including search areas prepared by Bob King for use by FSF in 2005. The figure compares the predicted and observed positions of the wrecks Finding Sydney Foundation (2005).

each of which has been added to the map. The errors for the two recommendations were 3 and 9 NM for *Kormoran* and *Sydney* respectively. Research and argument advanced subsequent to that date was superfluous, and served only to transfer responsibility for the success of the search.

Efficiency

The pink rectangle and the purple quadrant indicate the search areas recommended by the FSF for *Kormoran* and *Sydney* respectively, in 2005. The area of the pink rectangle, the recommended search box for *Kormoran*, reflected conventional statistical analysis based on the latitude and longitude values associated with each of the nine constraints. The area was therefore defined by the 95% confidence intervals for the x (longitude) and y (latitude) values based on positions attributed to each of the nine constraints. The area of the rectangle, 400 SNM, and the location of the wreck of *Kormoran*, can be compared with the area of 2200 SNM adopted by the FSF on the advice of Mearns a few weeks before the in-water search in 2008.

Explanatory power

The final cognitive analysis reflected the majority of the data summarized in **Table 1**. The cognitive solution was, furthermore, consistent with the known tracks of *Sydney* through the area, and the oceanographic solutions described above although these considerations did not contribute directly to the quantitative solution. The historical analyses exploited only the report originally extracted by Winter from Detmers' Diary (Winter, 1991), based on the noon position plus dead reckoning.

The research provided an accurate estimate of the position of the wreck of Kormoran, an efficient solution given a search box of < 400 SNM, and it exploited more than 50% of the items in the Kormoran Database. The research also provided an accurate estimate of the location of *Sydney*, based on a time series analysis of her reported bearing and distance from Kormoran over a 5 h period after the battle.

Opportunity Cost

The author did not review information about the 1968 search for the USS *Scorpion* prior to creation of the Kormoran Database. However, in 2006, when the FSF invited John Dunn and the author to table a new search proposal, we revisited the Search Definition problem, gave consideration to the Baysian description of the search for Scorpion (See Sontag and Drew, 1999), and tabled a new proposal that included provision for expert-based weighting for the individual constraints.

SKILL ACQUISITION

Table 2 is a summary of the recommendations advanced by the author and his colleagues between 1991 and 2005. The Search Definition problem was solved by Australian science for both Kormoran and Sydney by 2005.

The foregoing analysis described the collection and analysis of evidence concerning the location of the wreck of *Kormoran*. The improvement in performance summarized in **Table 2** and **Figure 9** does not reflect the performance of either a single individual or a regular team in the traditional sense of these terms. The task of wreck-hunting lies somewhere on a continuum of decomposability. At one extreme, the slow, and fundamental changes involved in landing safety on aircraft carriers (Wiegmann and Shappell, 2003, p. 5) and construction time for Liberty ships during World War II (Searle and Gody, 1945). Involved massively decomposable tasks where dozens or even hundreds of people contributed to the improvement in performance. The skills associated with accurate kicking for an oval-shaped Australian Football League football can be decomposed for learning purposes but they cannot be distributed across players or experts during a game. Each one has to kick the ball for himself or, on rare occasions, herself.

TABLE 2 | Summary of positions advanced by the author and his colleagues for Kormoran.

Sources	Coordinates	Error (NM)
STAGE 1: OCEANOGRAPHY/SAR		
Kirsner, 1991	25°58′S 111°24′E	22
Kirsner et al., 1992	26°01′S 111°16′E	12
	26°01′S 111°20′E	15
Kirsner and Hughes, 1993	~26°13′S 111°25′E	17
STAGE 2: COGNITION—CONVERGING OPERATIONS		
Kirsner, 1997a	26°15′S 111°E	10
Kirsner and Dunn, 1998a	26°15′S 111°E	10
Finding Sydney Foundation, 2001	~26°06′S 110°52′E	11
Finding Sydney Foundation, 2003	~26°10′S 111°10′E	7
STAGE 3: COGNITION—DECISION MODEL		
Kirsner and Dunn, 2004, 2008; Dunn and Kirsner, 2011; King, 2014; Kirsner and Dunn, 2014	26°04′S 111°02′E	3
Finding Sydney Foundation, 2005: Acknowledged Kirsner and Dunn	26°04′S 111°02′E	3
Finding Sydney Foundation, 2007: Acknowledged Kirsner and Dunn	26°04′S 111°02′E	3

Figure 1 identified a number of discipline-specific approaches to wreck-hunting, several of which were adopted by the author and his colleagues. The learning curve observed in **Figure 9** arguably reflects transitions across domains, from oceanography (1991–1993) to history to the cognitive sciences including adoption of a formal decision model. The critical drivers reflected: first, the construction of a substantial database; second, decisions about the viability of the report data; and, third, adoption of informal and then formal approaches to the use of multiple constraints. The research also reflected a coherent and evolving approach to a clearly defined problem concerning the location of the wreck of *Kormoran*, and, critically, it reflected input from a variety of disciplines, domains and scientists, individuals with diverse backgrounds.

Figure 9 includes the positions that the author and his colleagues tabled between 1991 and 2004. All of the positions in the plot are shown against an ordinate that indicates the distance between the position recommended and the position of the wreck of *Kormoran*, and the plot therefore reflects learning or skill acquisition. The reports formed three obvious groups involving

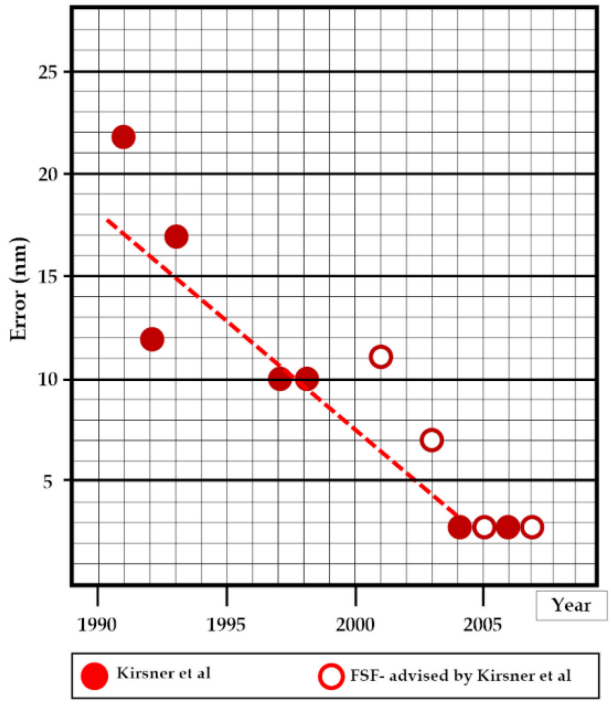


FIGURE 9 | Performance (Error) for the author and his colleagues between 1991 and 2005.

oceanography, informal analysis of the Kormoran Database, and formal or mathematical instantiation of the Minimum Distance Principle on the database respectively. The model tested a potentially infinite range of candidate locations, and selected the position that involved the smallest possible amount of movement for the set of nine constraints. The one and only solution associated with the third stage of the project therefore involved 26°04′S 111°02′E, a position just 2.7 nm from the wreck of *Kormoran*.

EXPERTISE

The path of improvement from 1991 to 2005 reflected input from no fewer than three disciplines, oceanography, history, and the cognitive sciences. One implication of this perspective is that the research involved both a *horizontal* trajectory, as we accepted and understood the limitations and opportunities associated with oceanographic and historical research respectively, and a *vertical* trajectory, as we implemented successive more and more powerful cognitive analyses of the survivors' reports. In so far as the project involved a *vertical* skill acquisition path, it conformed to the tradition established by John Anderson more than 30 years ago (e.g., Anderson, 1982), as well as the more specific benefits associated with transfer involving component process models, models that might or might not cross domain boundaries (e.g., Speelman and Kirsner, 2005).

The *horizontal* trajectory reflects an argument advanced by Engeström and his colleagues (e.g., Engeström and Sannino, 2010). According to Engeström (2014) for example,

> "*Learning by Expanding* challenges traditional theories that consider learning a process of acquisition and reorganization of cognitive structures within the closed boundaries of specific tasks or problems."

Elsewhere, Engeström (1996) proposed that learning is not restricted to "vertical movement across levels" but should also be viewed as "horizontal movement across borders." From a cognitive perspective however, the boundaries between the domains and the skills can be inherently fuzzy, and improvement will depend on comprehension and practice at the level of the component processes, and the discipline behind a given process might or might not be critical.

People acquire expertise or skill over a more or less unlimited range of domains and problems. The shear variety of the domains encompassed by human enterprise is formidable, and few attempts have been made to provide a *universal* model; that is, a model that covers all realms of human activity. To list but five disparate topics, a universal model would need to cater for the acquisition of skill or expertise in everything from cigar-rolling (Crossman, 1959) to survival in aerial combat (Spick, 1989), teamwork on the navigation bridge of a notional escort carrier (Hutchins, 1996), the reduction of flying accidents on Aircraft Carriers over 50 years (Wiegmann and Shappell, 2003) and construction times for Liberty ships (Searle and Gody, 1945). Wreck-hunting is just another cab off the ranks in the drive to describe and understand expertise and, if possible, define not only universal principles, such as the power law of learning, but a universal taxonomy as well.

Collins (2013) has offered a useful starting point in regard to a universal taxonomy with a three-dimensional model of expertise. The model was introduced under the heading of *Studies of Expertise and Experience,* and **Figure 10** honors the *Expertise-Space Diagram* depicted by that author.

The dimensions described by Collins were as follows:

1. The first or diagonal dimension is referred to as "Individual or group accomplishment" by Collins but the author has adopted a more traditional approach, treating this dimension as "Skill Acquisition," or, more simply, Skill, a term that usually pre-supposes qualitative changes in information processing strategies or processes as individuals or groups transit from novice to expert.
2. The second dimension described by Collins refers to the "transmission of domain-specific tacit knowledge," or Tacit Knowledge, involving either groups or individuals. The dimension is referred to as Tacit Knowledge in **Figure 10**, and depends on "immersion in the society of those who already possess it" (Collins, 2013, p. 3).
3. The third dimension referred to by Collins is "Esotericity," and this dimension is depicted on the vertical axis in Figure 12. According to Collins (2013, p. 5).

> "While traditional analyses take the word "expert" to refer only to rare, high-level, specialists, SEE (i.e., the model described by

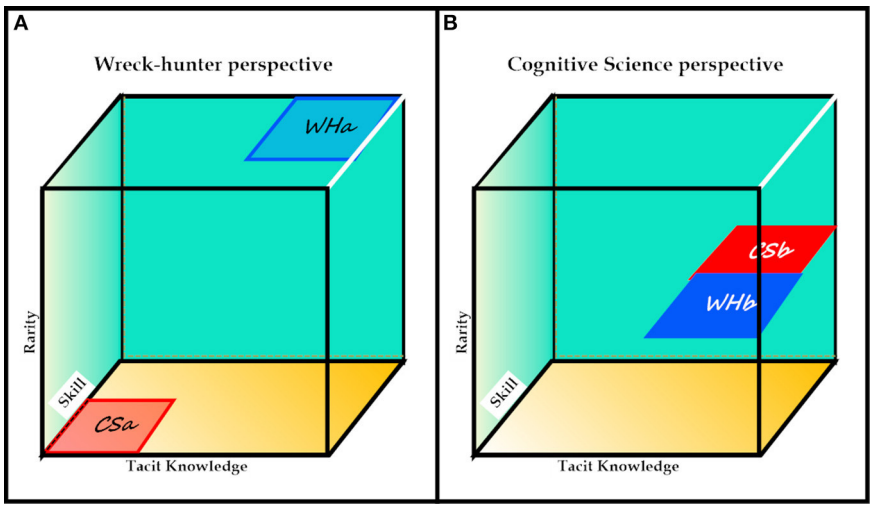


FIGURE 10 | Three dimensional model of expertise (from Collins, 2013). (A) reflects classification by a notional wreck-hunter. **(B)** reflects classification by the author, a classification that depends on the assumption that expertise for wreck-hunting is open to decomposition, a fourth dimension.

> Collins) considers that ordinary language-speaking, literacy and the like exhibit a high degree of expertise even though everyone has them—they are ubiquitous. This is, perhaps, one of the most radical contributions of SEE to the analysis of expertise as indicated by the initial strong opposition to the idea of "ubiquitous expertise" from philosophers and psychologists. Part of the task of this paper will be to try to make it obvious that the idea of ubiquitous expertise is a necessity if we are to avoid confusion."

In the following analysis, and in **Figure 10**, the term *rarity* is preferred to *esotericity* because of its frequency in natural language and its quantitative roots. A critical issue raised by Collins concerned the *rarity* of the relevant skills or expertise in his three-dimensional model of expertise. Collins questioned the traditional view that experts are of necessity "unusual individuals who have self-consciously devoted many hours of their lives to gaining a special ability." Instead, and based in part on the proposition that all native speakers of a language are experts to a greater or lesser extent in their native language, Collins mounted an argument against the esoteric or rarity characteristic of expertise, and proceeded to assert that "the idea of ubiquitous expertise is a necessity." Later, when faced with the challenge posed by racing car drivers, he proposed that the relevant skills form a body distinct from that of driving in general.

The importance of *decomposition* is evident. Changes in construction times for Liberty ships built in the US ranged from about 1.2 million man-hours per ship in the early days to less than 0.5 million man-hours per ship after 2000 or more vessels had been constructed. While a team of experts would have been essential to the design, co-ordination and management of each project, the improvement in ship-building times reflected many and widely distributed forms of expertise.

Another type of skill that reflects practice involved the performance and survival of fighter pilots in World Wars I and II (See Spick, 1989). Spick, for example, depicted the extent to which the probability of survival as a fighter pilot increased as a function of missions completed. Task analysis in this case involved a totally different picture from ship-building. The task of flying and fighting in World War I aircraft depended on indivisible expertise, expertise that accumulated with combat experience. The role of decomposition is quite different in this case however. While decomposition would have been possible and even desirable for instruction and training purposes, it was not possible to spread the skill across individuals under combat conditions, and each individual fighter pilot had to bring a full suite of skills to bear on the combat problem. Thus, while expertise can be distributed across thousands of engineers and craftsmen for ship-building, and reflect skill acquisition for the corporate entities as well as the individual tradesmen, a very different story applied to the performance of fighter pilots during World Wars I and II, and decomposition was not feasible under operational conditions.

But where does the foregoing analysis leave wreck-hunting in regard to expertise, or indeed, any research challenge that involves or could involve trans-disciplinary forms of expertise? The owners of the traditional forms of expertise might be reluctant to include provision for trans-disciplinary expertise, particularly if their background did not prepare them for challenges of this type.

Figure 10A depicts the model that a professional wreck-hunter might assert, and the model that was asserted or endorsed by virtually all of the parties involved the search for *Kormoran* and *Sydney*. However, as implied in the foregoing analysis, wreck-hunting can be treated as a decomposable example of expertise, involving a series of semi-independent skills or components. The critical issue implied by **Figure 10B** is that the set of reports from the Kormoran Database was open to analysis and interpretation by any one of a large number of cognitive scientists. In many cases we would have required support from historians and linguists but that can be assumed for multi- or trans-disciplinary projects. **Figure 10** therefore provides two frames of reference for a discussion of expertise in wreck-hunting; that of the wreck-hunter who claims that he or she is the only person who can solve the problem, and that of the cognitive scientist who claims that

wreck-hunting can be decomposed into component skills, skills that are widely distributed in the scientific community.

The argument outlined in this section has significant ramifications for the agencies and individuals responsible for *unprecedented* challenges such as those faced by the officials associated with the searches for *Kormoran* and *Sydney*, and, more recently, Malasia Airlines 370. The decision space should not be dominated by mate-ship and political expediency. Where inter- and trans-disciplinary opportunities are or might be relevant, effective leadership should involve scientifically informed and flexible leadership.

CONVERGING OPERATIONS, TRADING ZONES, AND "ENACTIVE" COGNITION

The author is not aware of any past attempts to consider or review wreck-hunting as a domain of expertise. The challenge is further complicated by the fact that it depends on several more specific forms of expertise, and few people will enjoy the complete set of skills involved. The project outlined in this paper therefore involved a de facto "trading zone" (See Thagard, 2005), or, to be more specific, an attempt to exchange ideas and approaches among navigators, oceanographers, historians, and cognitive scientists. The solution actually involved an expansion of triangulation, with nine as distinct from three Lines or Estimates of Position. However, the general principles guiding the cognitive approach to the challenge remained stable throughout the research, and relied on the presence of multiple constraints to negate the uncertainty and possible error associated with many if not all of the available reports. Given the central role of triangulation, a task that traditionally involved the use of maps, rulers, and Lines of Position, our analysis provides an interesting fit to the framework offered by Enactive Cognition (Froese et al., 2012). Specifically, it involves a task where the "cognitive agents" implement triangulation to solve a problem—to define the location of a wreck—and the physical vehicle for implementation, be it in a map, a head or a computer, is of secondary importance. Thus, triangulation constituted the critical scaffold for prediction, and the deep challenge facing us as scientists involved the selection and, if necessary, refinement of new Lines or Position. The solution was also consistent with earlier lines of argument involving: publications and papers describing the search and rescue solution, the SAR/Oceanography solution published by Sam Hughes, the Sunda Strait to Fremantle tracks taken by *Sydney* on earlier voyages, and the lifeboat tracks from the probable point of disembarkation from Kormoran to the coast.

As argued by Thagard (2005), science has changed out of recognition over the course of the twentieth century. Whereas, the early days of the century witnessed the establishment of the now traditional disciplines and divisions, some of which have been retained in the current curricula of universities, many of the critical advances in science and technology reflected migration to the boundaries of the established disciplines, as, like memes, they embarked on new inter-disciplinary journeys of their own. These transformations are particularly clear in the new and rapidly changing sciences, and the industries behind them, underwater target detection and forensic science being two obvious examples. It is also very clear in medical science and in medical training, where the nature and application of knowledge are undergoing similar transformations.

Much of the work described in this article reflected the author's origins in the cognitive sciences but it also capitalized on concepts, practices, and assumptions from older disciplines, involving oceanography and history in particular. Furthermore, and as argued by Thagard, trading zones are likely to flourish when they involve "people, places, organizations, ideas, and methods," and the arcane world of wreck-hunting provided a fascinating and challenging "trading zone."

CONCLUSION

The critical issue discussed in this article concerned the location of the wreck of the German raider *Kormoran* off the coast of Western Australia. An accurate solution required cognitive analysis of a chaotic database comprising more than 70 reports, a decision preceded by decisions to set aside oceanographic, navigation, map dowsing, historical and oral history arguments. The procedure reflected exceptional collaboration involving three or possibly more "trading zones," a critical step for innovation in science.

REFERENCES

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychol. Rev.* 89, 369–406.

Ballard, R. D. (1988). *The Discovery of the Titanic*. Toronto, ON: Madison Press.

Ballard, R. D. (1990). *The Discovery of the Bismarck*. London: Hodder & Stoughton.

Ballard, R. D. (2008). *Archaeological Oceanography*. Princeton, NJ; Oxford: Princeton University Press.

Ballard, R. D., and Archbold, R. (1999). *Return to Midway: The Quest to Find the Lost Ships from the Greatest Battle of the Pacific War*. Washington, DC: National Geographic Society.

Collins, H. (2013). Three dimensions of expertise. *Phenomenol. Cogn. Sci.* 12, 253–273. doi: 10.1007/s11097-011-9203-5

Crossman, E. R. F. W. (1959). A theory of the acquisition of speed-skill. *Ergonomics* 2, 153–166.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *Am. Psychol.* 34. 571–582.

Detmers, T. (1959). *The Raider Kormoran*. London: William Kimber.

Discenza, J. H., and Greer, J. W. (1994). *MELIAN II Search System*. Daniel H. Wagner Associates.

Dunn, J. C., and Kirsner, K. (2001). "Locating HMAS Sydney and HSK Kormoran by temporal triangulation," in *Report submitted to the Forum associated with the 60th Anniversary of the Loss of HMAS*, eds M. McCarthy (Sydney, NSW; Fremantle, WA: HMAS Sydney II Seminar).

Dunn, J. C., and Kirsner, K. (2011). The search for HMAS Sydney II: analysis and integration of survivor reports. *Appl. Cogn. Psychol. Appl. Cogn. Psychol.* 25, 513–527. doi: 10.1002/acp.1735

Engeström, Y. (1996). Developmental work research as educational research. *Nordisk Pedagogik* 16, 131–143.

Engeström, Y. (2014). *Learning by Expanding*. Cambridge, MA: Cambridge University Press.

Engeström, Y., and Sannino, A. (2010). Studies of expansive learning: foundations, findings and future challenges. *Educ. Res. Rev.* 5, 1–24. doi: 10.1016/j.edurev.2009.12.002

Finding Sydney Foundation (2001). *Problems and Opportunities for the search for HMAS*. Sydney, NSW: Powerpoint Presentation.

Finding Sydney Foundation (2003). *The Search for HMAS*. Sydney, NSW: Business Plan.

Finding Sydney Foundation (2005). *Search for HMAS Sydney II – 2005/6: Solving Australia's Most Enduring Maritime Mystery*. HMAS Sydney Search Pty Ltd. PowerPoint Presentation.

Finding Sydney Foundation (2007). *Status Report on Achievements and Opportunities*. Submission to the Commonwealth of Australia.

Finding Sydney Foundation (2008). *Final Report to the Department of the Environment, Water, Heritage and the Arts*. The Search for HMAS Sydney.

Froese, T., McGann, M., Bigge, W., Spiers, A., and Seth, A. K. (2012). The enactive torch: a new tool for the science of perception. *IEEE Trans. Haptics* 5, 365–375. doi: 10.1109/TOH.2011.57

Fry, H. (2012). *The M Room: Secret Listeners Who Bugged the Nazis in WW2*. London: Helen-Fry.

Gill, H. (1957, 1985). *Royal Australian Navy 1939-1942, Collins in Association with the Australian War Memorial*. Canberra, ACT: Australian War Memorial.

Green, J., McCarthy, M., and Penrose, J. (1984). Site inspection by remote sensing - the HMAS Sydney search: a case study. *Bull. Austr. Inst. Mar. Archaeol*. 8, 22–42.

Hansel, B. (2011). *Paddling Light: Lightweight Canoe and Kayak Travel*. Available online at: http://www.paddlinglight.com/articles/navigation-fixes-and-triangulation/

Hore, P., and Mearns, D. L. (2003). HMAS Sydney - an end to the controversy. *Naval Hist. Soc.* 4, 2–9.

Hughes, A. J. (1991). "A possible solution based on modern SAR Planning Techniques," in *The HMAS Sydney Forum*, eds M. McCarthy and K. Kirsner. Western Australia Maritime Museum Report #52. (Department of Maritime Archaeology).

Hutchins, E. (1996). *Cognition in the Wild*. Boston, MA: Massachusetts Institute of Technology.

Johnstone, D., Croigh, J., Cowan, J., and O'Driscoll, P. (2003). *Summary of HMAS SYDNEY (II) and KORMORAN – RAN Seapower Centre Lifeboat Workshop*.

King, B. (2014). "Commitment, Persistence and Science bring success," in *The Search for HMAS Sydney - An Australian Story*, eds E. Graham, B. Trotter, B. King, and K. Kirsner (Sydney, NSW: UNSW Press), 78–99.

Kaplan, I. T., and Carvellas, T. (1969). Response probabilities in verbal recall. *J. Verb. Learn. Verb. Behav.* 8, 344–349.

Kirsner, K. (1991). "HSK "Kormoran" versus HMAS "Sydney" - Converging operations in historical analysis," in *The HMAS Sydney Forum*, M. McCarthy and K. Kirsner Western Australia Maritime Museum Report #52. (Department of Maritime Archaeology).

Kirsner, K. (1997a). "Eagle in the Crow's Nest: cognitive analysis of a naval disaster," in *Paper presented to the Thirteenth Humanities Symposium, University of Western Australia, September*, 1997.

Kirsner, K. (1997b). *The War of the Ghosts: Using dusty records to hind-cast the locations of HMAS Sydney and HSK Kormoran*. Humanities Conference, UWA.

Kirsner, K., and Dunn, J. (1998a). "Feasibility of the search for HMAS Sydney and HSK Kormoran: oceanographic and cognitive issues," in *Submission to the Joint Standing Committee on Foreign Affairs, Defence and Trade Defence Sub-Committee: Inquiry into the Circumstances of the Sinking of HMAS Sydney*, Vol. 11, 2727–2742.

Kirsner, K., and Dunn, J. (1998b). "A review of the oceanographic and cognitive evidence that HSK Kormoran is at 28°39′S 113°22′E," in *Submission to the Joint Standing Committee on Foreign Affairs, Defence and Trade Defence Sub-Committee: Inquiry into the Circumstances of the Sinking of HMAS Sydney*, Vol. 16, 4023–4058.

Kirsner, K., and Dunn, J. (1998c). "Cognitive problems associated with the use of oral history data: supplement to 'Feasibility of the search for HMAS Sydney and HSK Kormoran: Oceanographic and Cognitive Issues'," in *Submission to the Joint Standing Committee on Foreign Affairs, Defence and Trade Defence Sub-Committee: Inquiry into the Circumstances of the Sinking of HMAS Sydney*, Vol. 18, 4311–4313.

Kirsner, K., and Dunn, J. (2004). *The Search for HSK KORMORAN and HMAS SYDNEY II: A Cognitive Perspective. Submitted to the FSF, the WAMM and David Mearns on December 1st, 2004. Published Subsequently on the UWA Site for Cognitive Analysis of Archival, Historical and Memory Databases*. Available onlie at: http://www.whereissydney.com/

Kirsner, K., and Dunn, J. C. (2008). *Search Definition in the Search for Kormoran and Sydney: Triumph for Cognitive Science. Submission requested by and submitted to the Cole Commission of Inquiry*.

Kirsner, K., and Dunn, J. C. (2014). "How Cognitive science put the FSF withing 'spitting distance' if the Kormoran wreck," in *The Search for HMAS Sydney - An Australian Story,* eds E. Graham, B. Trotter, B. King, and K. Kirsner (Sydney, NSW: UNSW Press), 126–153.

Kirsner, K., and Hughes, S. (1993). *HMAS Sydney and HSK Kormoran: Possible and Probable Search Areas*. Report No. 71. Fremantle, WA: Department of Maritime Archaeology, Western Australia Maritime Museum.

Kirsner, K., Hughes, S., Pearce, A., Penrose, J., Gauntlett, M., Steedman, R., et al. (1992). *The Search for HMAS Sydney and HSK Kormoran: A Proverbial Needle. Submitted to West Australian Museum*.

McDonald, D. (2003). *The Journey of Two Lifeboats. An Examination of the Voyages of KAPT LT Meyer and KAPT LT Von Malapert 19–25th Nov 1941*.

Mearns, D. (2007). *APPENDIX F. Summary of Search Plan Provided by Mr David Mearns, 4 July 2007: The Search for HMAS Sydney and HSK Kormoran. Submission from the Hon Bruce Billson, Minister for Veterans' Affairs, Minister Assisting Veterans Affairs, Federal Member for Dunkley, to The Hon John Howard Prime Minister, Parliament House*. Canberra, ACT: Finding Sydney Foundation.

Mearns, D. L. (1995). *Search for the Bulk Carrier Derbyshire: Unlicking the Mystery of the Bulk Carrier Shipping Disasters. Man-Made Objects on the Seafloor*. London: Society for Underwater Technology.

Mearns, D., and White, R. (2001). *Hood and Bismarck: The Deep-sea Discovery of an Epic Battle*. UK Channel 4 Books.

Miller, G. A., and Newman, E. B. (1958). Tests of a statistical explanation of the rank-frequency relation for words in written English. *Am. J. Psychol.* 71, 209–218.

Montgomery, M. (1981). *Who Sank the Sydney?* Sydney, NSW: Cassell.

Olson, W., Hore, P., Goldsmith, R., and Vickridge, G. (2001). *Archival Record*. Workshop Report. Perth, WA: HMAS Sydney Wreck Location Seminar.

Olson, W. (2000). *Bitter Victory*: *The Death of HMAS Sydney*. Perth: University of Western Australia Press.

Penrose, J. D., and Klaka, K. P. (1991). "Notes on the movement of wreck material from the area of the "Sydney" / "Kormoran" engagement," in *The HMAS Sydney Forum*, eds M. McCarthy and K. Kirsner Western Australia Maritime Museum Report #52. (Department of Maritime Archaeology).

Pearce, A. (1991). "Variability of ocean currents off shark bay," in *The HMAS Sydney Forum*, eds M. McCarthy and K. Kirsner. Western Australia Maritime Museum Report #52. (Department of Maritime Archaeology).

Searle, A. D., and Gody, C. S. (1945). Productivity changes in selected wartime shipbuilding programs. *Month. Labor Rev.* 61, 1132–1149.

Shannon, C. E. (1949). A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423, 623–656.

Sontag, S., and Drew, C. (1999). *Blind Man's Bluff*: *The Untold Story of American Submarine Espionage*. New York, NY: Perennial.

Speelman, C., and Kirsner, K. (2005). *Beyond the Learning Curve: The Construction of Mind*. Oxford: Oxford University Press.

Spick, M. (1989). *The Ace Factor*. Annapolis, MD: Airlife.

Steedman, R., and McCormack, P. (1991). "Backtracking the lifeboats and floats – a Metaocean view," in *The HMAS Sydney Forum*, eds M. McCarthy and K. Kirsner Western Australia Maritime Museum Report #52. (Department of Maritime Archaeology).

Tanner, W. P., and Swets, J. A. (1954). A decision-making theory of visual detection. *Psychol. Rev.* 61, 401–409.

Terkel, S. (1970). *Hard Times: An Oral History of the Great Depression*. New York, NY: New Press.

Thagard, P. (2005). "Being interdisciplinary: trading zones in cognitive science," in *Interdisciplinary Collaboration: An Emerging Cognitive Science*, eds S. J. Derry, C. D. Schunn, and M. A. Gernsbacher (Mahwah, NJ: Erlbaum), 317–339.

Wagenaar, W. A. (1986). My memory: a study of autobiographical memory over six years. *Cogn. Psychol.* 18, 225–252. doi: 10.1016/0010-0285(86)90013-7

Whittaker, W. (2000). *The Loss of HMAS Sydney - 1941: The Search for the Wreck of HSK Kormoran*. Unpublished Paper.

Wiegmann, D. A., and Shappell, S. A. (2003). *A Human Error Approach to Aviation Accident Analysis: The Human Factors Analysis and Classification System*. Aldershot, VT: Ashgate Publishing.

Winter, B. (1984). *HMAS Sydney: Fact, Fantasy and Fraud*. Brisbane, QLD: Boolerong Press.

Winter, B. (1991). "Loose Ends," in *Notes Submitted to the West Australian Maritime Museum for the Oceanography Workshop, November, 1991*.

Yeasmin, K. F. (2012). 'Triangulation' research method as the tool of social science research. *Bangl. Univ. Prof. J.* 1, 154–163. Available online at: http://www.bup.edu.bd/journal/154-163.pdf

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley; New York, NY: Hafner.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Kirsner. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 17 November 2015
doi: 10.3389/fpsyg.2015.01744

Attaining Automaticity in the Visual Numerosity Task is Not Automatic

Craig P. Speelman and Katrina L. Muller Townsend*

School of Psychology and Social Science, Edith Cowan University, Joondalup, WA, Australia

This experiment is a replication of experiments reported by Lassaline and Logan (1993) using the visual numerosity task. The aim was to replicate the transition from controlled to automatic processing reported by Lassaline and Logan (1993), and to examine the extent to which this result, reported with average group results, can be observed in the results of individuals within a group. The group results in this experiment did replicate those reported by Lassaline and Logan (1993); however, one half of the sample did not attain automaticity with the task, and one-third did not exhibit a transition from controlled to automatic processing. These results raise questions about the pervasiveness of automaticity, and the interpretation of group means when examining cognitive processes.

Keywords: automaticity, skill acquisition, average, individual differences, practice

OPEN ACCESS

Edited by:
Pietro Cipresso, IRCCS Istituto Auxologico Italiano, Italy

Reviewed by:
Evgueni Borokhovski, Concordia University, Canada
Konrad Schnabel, International Psychoanalytic University Berlin, Germany

****Correspondence:***
Craig P. Speelman
c.speelman@ecu.edu.au

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *24 June 2015*
Accepted: *30 October 2015*
Published: *17 November 2015*

Citation:
Speelman CP and Muller Townsend KL (2015) Attaining Automaticity in the Visual Numerosity Task is Not Automatic. Front. Psychol. 6:1744. doi: 10.3389/fpsyg.2015.01744

INTRODUCTION

Speelman and McGann's (2013) paper contained a clear message for Psychology: Be wary of phenomena that are discovered on the basis of average group data. Speelman and McGann (2013) argued that the averaging process typically used in the analysis methods adopted by Psychology can mask individual results that may actually be counter to those revealed by the group results. As a result, group results may not be an accurate reflection of the behavior of many, and possibly most, individuals in the group.

Most of the phenomena we teach as the basic facts in an introductory course in Cognitive Psychology have typically been generated by experiments where groups of subjects perform the same task under various conditions. The classic result is usually observed and interpreted as a pattern of differences between group and/or condition means. That is, the take-home message from these experiments is usually represented as a pattern of results that are generated by averaging across the results of individuals. This results in a 'clean' picture of behavior where the noise associated with individual differences has effectively been removed. Thus well-known effects such as the word superiority effect, the serial position curve, the power law of learning, and the phonological similarity effect have been well replicated by different researchers and under different conditions, but they all are observed by averaging data collected from groups of individuals. Speelman and McGann (2013) argued that such effects may not be as pervasive as their replicability suggests. That is, although the effects can be replicated easily enough, they may only exist when the data from several individuals are combined, and as a result may not reflect the cognitive processes of many, and at worst, any individuals in that group.

Speelman and McGann (2013) reported results from a replication of the Word Superiority Effect. Although the average performance of the sample in their experiment replicated the classic effect, an examination of the performance of individuals in the sample revealed that very few people produced results consistent with the effect. Speelman and McGann (2013) argued that such a result

should reduce the confidence we have in using means to reveal information about fundamental cognitive processes. A further implication of this result is that it may be prudent to determine the extent to which individuals demonstrate performance patterns that have to date been demonstrated in group results.

The research reported in this paper was designed to examine a well-replicated finding in the field of attention. From at least as far back as the 1970s, researchers (e.g., Posner and Snyder, 1975; Schneider and Shiffrin, 1977; Hasher and Zacks, 1979) have drawn a distinction between automatic and controlled/conscious/effortful forms of mental processing. Controlled processes are typically exhibited early in the practice of a task, while we are more likely to perform automatically after a long period of practice. Controlled, deliberate psychological processes are used for difficult and unfamiliar tasks. These processes operate serially, use substantial cognitive resources, require attention, and are flexible. In contrast, automatic processes are used for easy and familiar tasks, operate in parallel, require very few cognitive resources, do not require attention, and are difficult to modify (Speelman and Maybery, 1998). Automatic performance only comes after extensive practice. Thus, with sufficient practice under appropriate conditions (Schneider and Shiffrin, 1977; Shiffrin and Schneider, 1977) one can develop the ability to respond in an automatic fashion to particular stimuli.

Most theories of cognitive skill acquisition describe mechanisms by which practice produces a shift from controlled to automatic processing (e.g., Logan, 1988; Anderson and Lebiere, 1998). Although these theories propose different means by which practice leads to more efficient processing, all of the theories lead to the same prediction: with sufficient practice of a task where the stimulus–response relationship is consistent, performance will reach the stage where perception of a known stimulus will trigger an automatic response (i.e., seeing '$3 \times 4 = ?$' will automatically lead to a response of '12').

This view of the development of automatic processing has influenced ideas of how we acquire complex skills. When we initially embark on the acquisition of such skills, effort, and attention are focused on basic, low level tasks (e.g., recognizing letters when learning to read). These tasks are practiced until processes are developed that perform this task automatically. Initially, these processes require most of the available cognitive resources to proceed. Little capacity is available for any other task (e.g., reading words). Once these processes have become automatic, however, sufficient cognitive resources are available for the person to attempt higher level tasks (e.g., reading words). Importantly, higher level tasks (i.e., reading words) are considered to operate on the outcomes from lower level tasks (e.g., letter identification; Karmiloff-Smith, 1979). With further practice, processes will be developed that are specific to the higher level task and these in turn may become automatic, enabling further developments in the level of skill (Speelman and Kirsner, 2005). This view is clearly articulated in mainstream educational practice (e.g., Cumming and Elkins, 1999; Caron, 2007; Baroody et al., 2009).

Automaticity can be attained quickly with simple tasks. Lassaline and Logan (1993) trained subjects on a visual numerosity task. In this task pictures of dots or similar, ranging in number from 6 to 11 and arranged in a seemingly random manner, are presented on a computer screen, one at a time (**Figure 1**). Subjects are required to indicate how many dots are presented, as quickly as possible. Typically, the speed with which subjects can perform this task is associated with the number of dots on the screen. That is, the more dots in a picture, the slower the reaction time (RT). However, when pictures are repeated, and subjects have a lot of practice at the task, eventually their RTs are no longer associated with the number of dots in a picture – subjects respond to each picture with equivalent speed (**Figure 2**). According to Logan's (1988) theory of skill acquisition, early in training subjects count the dots and this typically takes longer to complete the more dots there are in a picture. Late in training, subjects are more likely to recognize pictures and so remember the number of dots rather than have to count them. As a result they can respond to each picture with the same speed and hence RT will not be a function of the number of dots in a picture. An RT line with a zero slope, therefore, indicates automaticity of this response.

In Lassaline and Logan's (1993) experiments, subjects reached this state after four sessions of training (1920 trials and 64 repetitions per item). Lassaline and Logan (1993) interpreted this change in the pattern of RTs as subjects moving from a counting strategy early in practice (a controlled process) to a memory strategy later in practice (i.e., subjects recognized each picture and remembered the correct response – an automatic process).

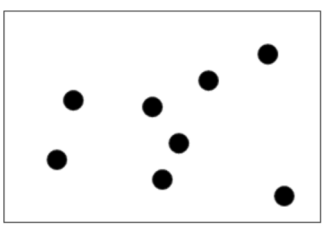


FIGURE 1 | An example of the type of dot picture used by Lassaline and Logan (1993). Subjects are asked to indicate the number of dots in the picture.

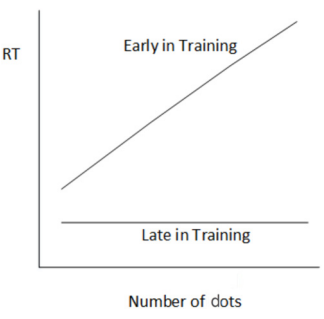


FIGURE 2 | Reaction time (RT) in the visual numerosity task as a function of number of dots in each stimulus picture.

A similar explanation invoking a transition from controlled to automatic processing has been used to account for results in the alphabet arithmetic task (Logan, 1988) and memory scanning (Schneider and Shiffrin, 1977). The fact that the pattern of performance changes that accompany practice are so easily replicated has no doubt fostered confidence in this explanation. What is not clear in any of this research, however, is the extent to which this transition occurs in individuals. The traditional approach in this area of research, as in many other areas of cognitive psychology, is to collect data from groups of subjects, and analyze the trends in the average results. Certainly, theories such as Logan's (1988) have been developed to explain the average results. But are these theories a good explanation for what occurs in the minds of all individuals when they practice any of these tasks? It is difficult to answer this question because it is not traditional practice to publish data that reflects how well the average trends match the pattern of results produced by each individual. The aim of the experiment reported here was to provide data that could be used to answer this question with respect to the visual numerosity task.

The experiment was an attempt to replicate the results reported by Lassaline and Logan (1993) and depicted in **Figure 2**. In addition, we looked at the individual RT results of each subject to determine the extent to which the apparent transition from controlled to automatic processing occurs in a sample of people. If a result similar to that reported by Speelman and McGann (2013) was obtained – that is, that the group results replicate the classic effect but that a substantial proportion of the sample do not show the effect – then this would raise questions regarding the validity of theories designed to explain the group results.

MATERIALS AND METHODS

Subjects

Eighteen psychology students from Edith Cowan University voluntarily participated in the study. The inclusion criteria required participants to have 'corrected' or 'corrected-to-normal' vision and English as their primary language. The participants' ages ranged from 19 to 65 years (**Table 1**). Participants were reimbursed with a $20 shopping voucher for their time. This experiment was approved by the Edith Cowan University Human Research Ethics Committee. All subjects granted their written informed consent to participate in the experiment.

Design and Stimuli

The visual numerosity task used in this experiment was performed as part of a larger task used to examine transfer of training issues. Each trial had two parts. In the first part of each trial, a configuration of asterisks was presented on the computer screen. Subjects were asked to indicate the number of asterisks as quickly as possible by pressing one of six buttons on a response box. The second part of each trial involved subjects adding a number presented on the screen to the number of dots that had been presented in the previous part. Subjects then decided whether the sum was an odd or even number, indicating their decision by pressing the appropriate button on the response

TABLE 1 | Slopes of regression lines (ms/asterisk) fitted to RT data as a function of numerosity for each subject.

Participant	Age (years)	Early slope	Mid slope	Late slope	Matches auto pattern	Auto (<100 ms)
1	48	157.83 (0.42)	264.35 (0.71)	329.44 (0.66)		
2	55	221.06 (0.74)	174.39 (0.64)	147.55 (0.52)	y	
3	47	194.95 (0.21)	−21.02 (0.01)	−45.10 (0.02)	y	y
4	49	106.44 (0.23)	−16.02 (0.00)	−77.38 (0.15)	y	y
5	49	425.72 (0.94)	335.25 (0.75)	377.09 (0.80)		
6	57	247.66 (0.82)	224.68 (0.69)	263.65 (0.69)		
7	25	336.01 (0.68)	−95.78 (0.07)	−71.43 (0.07)	y	y
8	23	381.07 (0.81)	305.30 (0.81)	218.92 (0.92)	y	
9	27	264.21 (0.90)	361.59 (0.91)	250.73 (0.54)		
10	31	318.85 (0.84)	41.37 (0.09)	6.13 (0.00)	y	y
11	28	193.07 (0.38)	−15.49 (0.00)	−126.67 (0.07)	y	y
12	39	277.37 (0.85)	90.37 (0.27)	59.12 (0.08)	y	y
13	65	367.46 (0.50)	254.99 (0.32)	248.84 (0.55)	y	
14	20	309.01 (0.88)	537.37 (0.52)	232.21 (0.34)		
15	23	625.63 (0.58)	86.52 (0.13)	−54.47 (0.25)	y	y
16	19	447.66 (0.70)	551.07 (0.84)	450.31 (0.78)		
17	30	290.94 (0.83)	108.88 (0.24)	−11.66 (0.00)	y	y
18	48	216.73 (0.52)	38.69 (0.01)	−65.27 (0.12)	y	y
Mean/total	37.94	298.98 (0.94)	179.25 (0.57)	118.44 (0.46)	12/18	9/18
r with age		-0.42^{ns}	-0.19^{ns}	0.09^{ns}		

Values in parentheses are r^2 values for the regression lines. Matches auto pattern: y = yes, slopes descend from Early to Mid to Late. Auto (<100 ms): y = yes, slope = 100ms or less by the Late period. ns = $p > 0.05$.

box. Only data from the visual numerosity part of each trial is considered in this paper.

Six stimuli were prepared for this experiment, one for each level of numerosity from 6 to 11. In each stimulus, asterisks were arranged in a pseudo-random manner, with the constraint that each asterisk was separated from other asterisks by at least 1 cm.

Procedure

Subjects were provided with 12 trials of practice using stimuli that were not used in the experimental trials but which were similar in appearance to the experimental stimuli. Once participants fully understood the procedure the experimental trials began. In part one of each trial a fixation point appeared in the centre of the screen for 250 ms, followed by a configuration of asterisks ranging in number from 6 to 11. Subjects were required to determine the number of asterisks. The picture remained on screen until a response was made on the response box by pressing one of the keys labeled 6–11. A blank screen then followed for 250 ms. Participants were then asked to add a number to the number of stars just identified and determine whether the answer was odd or even by pressing the corresponding keys. Participants were instructed to respond as accurately and as quickly as possible. Trials were presented in blocks of six. The six trials within each block were presented in a random order. There were 50 blocks of trials, leading to 300 trials, with each stimulus being presented 50 times.

RESULTS

Accuracy of responses was examined to ensure that participants were not guessing with their responses. All participants maintained a mean accuracy above 80% for each block of trials.

Average RT across all levels of numerosity for each block was calculated. **Figure 3** shows RT as a function of practice for each block. Mean RT for the group became faster over the experiment, and is well described by a power function.

Blocks were examined in phases (each block consisted of six trials): Early (blocks 1–10); Mid (blocks 21–30); and Late (blocks 41–50). Mean RT for each level of numerosity for each phase is presented in **Figure 4**. The slope of a regression line relating response latency to numerosity was calculated for each of the three phases to determine whether automaticity was reached. These values are presented in **Table 1**. Although the slope values follow the pattern of results reported by Lassaline and Logan (1993) – that is, the slopes decline in value from Early to Late in practice – the Late result does not reach 0 ms/asterisk, as would be expected if the results reflected complete automaticity. The slope value in the Late phase (118.44 ms/asterisk), however, is consistent with the slope value reported by Lassaline and Logan (1993) in several experiments after a similar amount of practice (circa 100 ms/asterisk).

Regression lines were also fitted to the RT data as a function of numerosity for each phase and for each subject. The slope values for these lines are presented in **Table 1**. Two analyses were performed with this data. The first examined the number of subjects whose slope values followed the pattern of the slopes

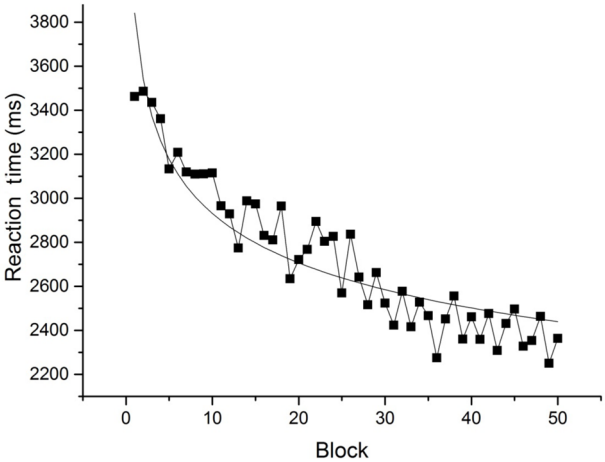


FIGURE 3 | Mean RT for each block of trials. The smooth line is the best-fit power function (RT = 342.30 + 3500 × block$^{-0.13}$, $r^2 = 0.86$, and rmsd = 99.68).

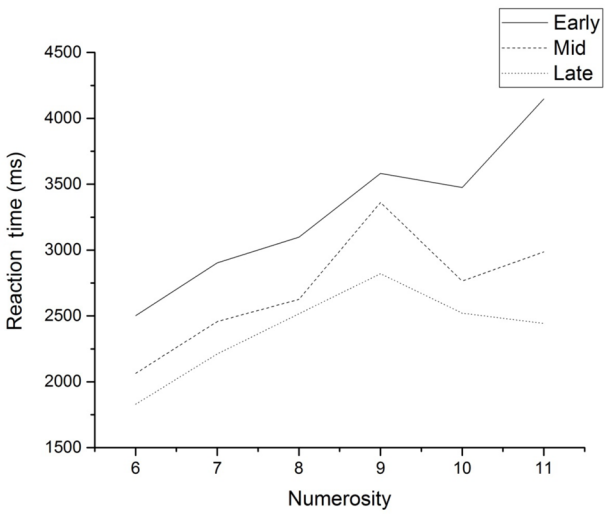


FIGURE 4 | Mean RT as a function of numerosity for the three phases of the experiment.

calculated on group data. Twelve out of 18 (67%) subjects fell into this category, although subject 13 barely shows a reduction in slope value from the Mid to the Late phase. The other analysis of slope values looked at whether or not a subject reached a slope value of 100 ms/asterisk or less, to signify whether a subject had reached automaticity. This was the same cut-off point as used with the group results. The slope values of nine out of 18 (50%) subjects met this criterion.

A further test was performed to determine whether the pattern of slope value changes from phase to phase was consistent amongst the subjects. Kendall's coefficient of concordance indicated that there was some degree of consistency amongst subjects in their slope value changes ($W = 0.373$, $X^2(2) = 13.444$, $p = 0.001$). However, the fact that W was not equal to 1 indicates

that there was not complete unanimity in the pattern of slope changes, and being closer to 0 than 1 supports the fact that there was a sub-set of subjects that did not show the typical pattern of slope reductions throughout training.

To explore whether some characteristic of the subjects was associated with the likelihood of them attaining automaticity, Pearson correlation coefficients were calculated between subject's age and the regression slopes in the three practice phases. These values are reported in **Table 1**. None of these correlations were statistically significant.

DISCUSSION

This experiment replicated the result reported by Lassaline and Logan (1993). That is, practice with the visual numerosity task resulted in a change in the pattern of performance, with RT early in practice being a function of numerosity (i.e., the more asterisks to be counted, the longer the RT), whereas later in practice the relationship between RT and numerosity became weaker. These results can, therefore, be explained by the typical account, that performance has moved from a controlled form of processing early in practice (i.e., counting asterisks in a serial manner) to automatic processing (i.e., subjects recognize each stimulus and remember the number of asterisks in the picture). At least, this is what the group results suggest.

A different picture is apparent when the results of individual subjects are considered. First, only half of the sample produced results that suggested they had reached automaticity with the task. Second, at least one-third of the sample did not show results consistent with the group trend that replicated Lassaline and Logan's (1993) result. Thus, for this latter sub-group, there is no evidence that their results reflected a transition from controlled to automatic processing. So, although the overall group results reflect a pattern that describes well the results of two thirds of the sample, they do not reflect the pattern of behavior in all subjects. Indeed, a sizeable minority exhibited results that suggest there was no move toward automatic performance with the visual numerosity task.

One possible explanation for why so many people in this experiment did not attain automatic performance concerns the nature of the task used in this experiment, which differed from that used by Lassaline and Logan (1993). In this experiment there were two parts to each trial. The results reported in this paper only concerned the first part of each trial, the part that matched the task used by Lassaline and Logan (1993). It is possible that the presence of the second part of each trial in this experiment may have contributed to many subjects not showing a transition to automatic performance. On the other hand, the fact that the group results for this experiment were consistent with the group results reported by Lassaline and Logan (1993) indicates that the two-part structure to each trial did not affect the overall results. It is therefore not possible to rule this explanation in or out at this stage without knowing the individual results of subjects in the Lassaline and Logan (1993) experiments. It is worth noting, though, that in other visual numerosity experiments we have conducted in our laboratory with a similar two-part structure to each trial, the group results suggested a transition from controlled to automatic processing, whereas the individual results indicated that this transition was not universal. That is, in three experiments, 7/16, 15/20, and 23/40 people showed a transition to automaticity.

Another possible point of difference between our experiment and those reported by Lassaline and Logan (1993) concerns age. Lassaline and Logan (1993) did not report the ages of their subjects, only that they were undergraduate Psychology students. Our subjects also were undergraduate Psychology students, however, given the age profile of students at Edith Cowan University, the age range of our subjects being 19–65 years is not unusual. It may well be the case that the age range of our subjects is larger than the age range of subjects in Lassaline and Logan's (1993) experiments; however, age was not correlated with our measure of automaticity (regression line slopes) at any point in the experiment, and so cannot explain why so many subjects did not attain automaticity.

At the least, this experiment questions the conclusions that can be drawn from group results on the visual numerosity task. Although the group results are consistent with a transition from controlled to automatic processing, they do not reflect the performance of all subjects in the group. Even though there was a similar number of repetitions per item in this experiment (50 repetitions/item) to that in Lassaline and Logan's (1993) experiments (up to 64 repetitions/item), and so a similar opportunity to attain automaticity, some subjects in this experiment showed no evidence of moving toward automatic processing. Indeed, it seems that these people continued to count asterisks throughout the experiment. Thus, the transition from controlled to automatic processing as a result of practice with a task, which is a feature of many theories of skill acquisition (e.g., Logan, 1988; Anderson and Lebiere, 1998), may not be an automatic feature of skill acquisition, at least for some people. Other work has demonstrated that not all experimental subjects adopt more efficient performance strategies when acquiring skills, but rather just improve the application of a less-efficient strategy (Rowell et al., 2015). It is now an open question as to why some people do not exhibit this transition when many others do. Importantly, this is a question that would never arise without attention to differences between individual and group results (Speelman and McGann, 2013).

REFERENCES

Anderson, J. R., and Lebiere, C. (eds). (1998). *Atomic Components of Thought*. Mahwah, NJ: Erlbaum.

Baroody, A. J., Bajwa, N. P., and Eiland, M. (2009). Why can't Johnny remember the basic facts? *Dev. Disabil. Res. Rev.* 15, 69–79. doi: 10.1002/ddrr.45

Caron, T. A. (2007). Learning multiplication: the easy way. *Clear. House J. Educ. Strat. Issues Ideas* 80, 278–282. doi: 10.3200/TCHS.80.6.278-282

Cumming, J. J., and Elkins, J. (1999). Lack of automaticity in the basic addition facts as a characteristic of arithmetic learning problems and instructional needs. *Math. Cogn.* 5, 149–180. doi: 10.1080/135467999387289

Hasher, L., and Zacks, R. T. (1979). Automatic and effortful processes in memory. *J. Exp. Psychol. Gen.* 108, 356–388. doi: 10.1037/0096-3445.108.3.356

Karmiloff-Smith, A. (1979). Micro- and macrodevelopmental changes in language acquisition and other representational systems. *Cogn. Sci.* 3, 91–117. doi: 10.1207/s15516709cog0302_1

Lassaline, M. E., and Logan, G. D. (1993). Memory-based automaticity in the discrimination of visual numerosity. *J. Exp. Psychol. Learn. Mem. Cogn.* 19, 561–581. doi: 10.1037/0278-7393.19.3.561

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychol. Rev.* 95, 492–527. doi: 10.1037/0033-295X.95.4.492

Posner, M. I., and Snyder, C. R. R. (1975). "Attention and cognitive control," in *Proceedings of the Loyola Symposium on the Information Processing and Cognition*, ed. R. L. Solso (Hillsdale, NJ: Erlbaum), 55–85.

Rowell, N. E., Green, J. J., Kaye, H., and Naish, P. (2015). Information reduction – more than meets the eye? *J. Cogn. Psychol.* 27, 89–113. doi: 10.1080/20445911.2014.985300

Schneider, W., and Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychol. Rev.* 84, 1–66. doi: 10.1037/0033-295X.84.1.1

Shiffrin, R. M., and Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychol. Rev.* 84, 127–190. doi: 10.1037/0033-295X.84.2.127

Speelman, C. P., and Kirsner, K. (2005). *Beyond the Learning Curve: The Construction of Mind*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780198570417.001.0001

Speelman, C. P., and Maybery, M. (1998). "Automaticity and skill acquisition," in *Implicit and Explicit Mental Processing*, eds K. Kirsner, C. P. Speelman, M. Maybery, A. O'Brien-Malone, M. Anderson and C. MacLeod (Hillsdale, NJ: Erlbaum).

Speelman, C. P., and McGann, M. (2013). How mean is the mean? *Front. Psychol.* 4:451. doi: 10.3389/fpsyg.2013. 00451

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Speelman and Muller Townsend. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 07 August 2015
doi: 10.3389/fpsyg.2015.01171

To transform or not to transform: using generalized linear mixed models to analyse reaction time data

*Steson Lo * and Sally Andrews*

School of Psychology, University of Sydney, Sydney, NSW, Australia

Linear mixed-effect models (LMMs) are being increasingly widely used in psychology to analyse multi-level research designs. This feature allows LMMs to address some of the problems identified by Speelman and McGann (2013) about the use of mean data, because they do not average across individual responses. However, recent guidelines for using LMM to analyse skewed reaction time (RT) data collected in many cognitive psychological studies recommend the application of non-linear transformations to satisfy assumptions of normality. Uncritical adoption of this recommendation has important theoretical implications which can yield misleading conclusions. For example, Balota et al. (2013) showed that analyses of raw RT produced additive effects of word frequency and stimulus quality on word identification, which conflicted with the interactive effects observed in analyses of transformed RT. Generalized linear mixed-effect models (GLMM) provide a solution to this problem by satisfying normality assumptions without the need for transformation. This allows differences between individuals to be properly assessed, using the metric most appropriate to the researcher's theoretical context. We outline the major theoretical decisions involved in specifying a GLMM, and illustrate them by reanalysing Balota et al.'s datasets. We then consider the broader benefits of using GLMM to investigate individual differences.

Keywords: RT transformations, generalized linear mixed-effect models, mental chronometry, interaction effects, additive factors

OPEN ACCESS

Edited by:
Craig Speelman,
Edith Cowan University, Australia

Reviewed by:
Michael Smithson,
Australian National University, Australia
Guillermo Campitelli,
Edith Cowan University, Australia

****Correspondence:***
Steson Lo,
School of Psychology, University of Sydney, Griffith Taylor Building (A19), Sydney, NSW 2006, Australia
steson.lo@sydney.edu.au

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *20 April 2015*
Accepted: *24 July 2015*
Published: *07 August 2015*

Citation:
Lo S and Andrews S (2015) To transform or not to transform: using generalized linear mixed models to analyse reaction time data. Front. Psychol. 6:1171. doi: 10.3389/fpsyg.2015.01171

Introduction

A central theme of this special issue is how the uncritical use of statistical procedures in psychological research can lead researchers to draw incorrect theoretical and practical conclusions. From a procedure as simple as averaging over a set of data points, Speelman and McGann (2013) elaborated how the resulting value is often used to draw conclusions that violate many theoretical positions describing individual, or even moment to moment, volatility in human cognitive systems.

Similarly, Trafimow (2014) expressed concern over the use of statistical techniques like related-samples *t*-tests, which appropriately assess differences between individuals (e.g., do changes in attitudes differ across people on average because of variable X), but are ubiquitously used inappropriately to address hypotheses formulated within each individual (e.g., does variable X cause a particular person's attitude to differ).

Extending this theme, we focus on another simple procedure that can lead researchers to draw misleading theoretical conclusions if applied uncritically: the routine transformation of the dependent variable to meet assumptions of normality in inferential statistics. In particular, we

address issues associated with analysis of reaction time (RT) data—one of the most commonly used dependent variables in cognitive psychological research.

For over 100 years, cognitive psychologists have used RT to investigate unobservable mental processes (Donders, 1868/1969; Luce, 1986). These investigations are based on two fundamental assumptions: (i) mental processes take time to complete, and that (ii) each measured RT reflects a composite of several distinct stages of processing (e.g., visual encoding, mental processing, and response selection). This "chronometric" approach to mental processes underpins many paradigms in cognitive psychological research (Posner, 1978).

Because any single RT might contain idiosyncratic processes, such as lapses in attention, orthogonal to the mental process under investigation (however see Speelman and McGann, 2013 for an alternative perspective), researchers usually recruit multiple participants and subject them to multiple measurements of RT. This distribution of RTs obtained in simple decision tasks is invariably positively skewed. In traditional mean-based ANOVA analyses, issues regarding skew are typically ignored because the method has been repeatedly shown to be "robust to violations of normality" (e.g., Glass et al., 1972; Harwell et al., 1992; Lix et al., 1996). Consequently, many cognitive theories have been developed and validated against such mean RT data, raising many of the interpretive problems highlighted by Speelman and McGann (2013).

In response to such theoretical limitations, there have been two major developments in analysis of RT in cognitive research relevant to the themes of this issue. First, many researchers have moved "beyond mean RT" (Balota and Yap, 2011) by analysing changes in the RT distribution at a more fine-grained level in order to yield more accurate measures of group performance (Heathcote et al., 2004). Application of these procedures has allowed researchers to conduct sophisticated tests of cognitive theories that cannot be distinguished on the basis of mean RT alone (e.g., Heathcote et al., 1991; Andrews and Heathcote, 2001; Yap et al., 2009). For example, Yap et al. (2009) reported that an individual's vocabulary level modulated how word frequency and semantic priming affected the shape of their RT distribution. They found additive effects between these factors across the RT distribution for those of high vocabulary, suggesting that semantic priming was automatically triggered for both high and low frequency words among these people with highly fluent lexical representations. In contrast, those of low vocabulary showed interactive effects, particularly for slow responses, suggesting that the increased skew associated with greater priming for less familiar, low frequency target words might be due to strategic use of semantic information. Analyses of individual RT distributions have therefore proved to be useful in identifying and interpreting individual differences in speeded response tasks.

A second recent response to limitations of traditional ANOVA analyses of mean RT, which is the focus of the present paper, is the use of linear mixed-effect models (LMMs). LMMs have become increasingly prevalent within many areas of science, because they are able to account for random populations that share a nested relationship like hospitals chosen from different districts (Carey, 2002), or blocked relationships like fertilizer treatment on samples over different soil plots (Lane, 2002). Within cognitive psychology, LMMs have had the strongest recent impact in psycholinguistics, because the use of mean RT in traditional ANOVA analyses has been unable to capture the crossed relationship between counterbalanced sets of linguistic stimuli presented to different subjects (Clark, 1973; Forster and Dickinson, 1976; Baayen, 2008). LMMs provide a statistical solution to this problem (Baayen et al., 2008), and have become the recommended form of analysis in high impact journals within the field.

Importantly, LMMs have the potential to address many of the problems raised by Speelman and McGann (2013) about the use of mean RT, because the ability of these models to simulate the multi-level structure of the designs described above eliminates the need to average data across subjects, items, plots, or hospitals. This crucial property of LMMs therefore provides a powerful and refined method for investigating interactions of experimental effects with individual and item differences that cannot be investigated in traditional ANOVA approaches because they do not collapse across these variables. For example, by exploring the variance/covariance parameters, Kliegl et al. (2010) showed that individuals who responded more quickly tended to produce larger masked repetition priming effects in a lexical decision task. Across individual trials, Kinoshita et al. (2011) showed that sensitivity to the difficulty of the previous trial interacted significantly according the prime-target relationship and task environment in a parity judgment task. Thus, LMMs have the potential to accommodate the different levels of analysis required to "optimize both scientific rigor and sensitivity to individual variability" that was identified as one of the goals outlined in this Special Issue.

Although the sophistication of LMMs present a significant leap forward for individual differences research, their application is complicated for skewed dependent variables like RT because current guidelines for LMM recommend that researchers transform their RT for two reasons. The first is that skewed RT data can affect the estimate of the mean, thus distorting the outcome of statistical tests. For example, Baayen (2008) recommends transforming RT data to avoid a situation in which "just a few extreme outliers might dominate the outcome, partially or even completely obscuring the main trends characterizing the majority of datapoints" (p. 33). The second reason is that non-normally distributed residuals produced by skewed data reflect a non-constant heteroscedastic pattern that affects the precision with which the standard error of the mean is estimated (Cohen et al., 2003). Therefore, researchers are expected to use the Box–Cox procedure (Box and Cox, 1964) to identify a transformation that allows them to meet the Gaussian assumptions of normality and homoscedasticity. For RTs, the transformation that best satisfies this mathematical assumption is often the reciprocal or inverse RT (Balota et al., 2013).

To Transform or Not to Transform?

Unfortunately, routinely applying such transformations has important theoretical implications. For example, applying a non-linear (e.g., log, inverse) transformation to the dependent variable

not only normalizes the residuals, but also distorts the ratio scale properties of measured variables, such as dollars, weight or time (Stevens, 1946). As a concrete example within the aging literature, two samples—one older and one younger—might exhibit differential benefits in RT when the preceding prime word was semantically related to the target (e.g., nurse–doctor) relative to when it was semantically unrelated (e.g., plane–doctor) (e.g., 600 and 700 ms for the younger adults, and 780 and 910 ms for the older adults). However, on the log-transformed scale, differences between these two samples are obscured because on this scale the differences disappear [e.g., log(700 ms)−log(600 ms) = 0.15415; log(910 ms)−log(780 ms) = 0.15415] (i.e., there is no interaction between age and priming).

While many readers will recognize these discrepant results as another example of "scale dependent" interactions (Loftus, 1978), the critical question that we wish to address is what the correct scale should be in "chronometric" research. According to the "mental chronometry" approach (Posner, 1978), the answer is clearly raw RT. Differences in RT over experimental conditions are assumed to directly reflect differences in the amount of time taken to perform these mental operations (Townsend, 1992). In the example above, additive effects suggest that automatic spreading activation, which is thought to underlie semantic priming, proceeds in much the same way for both younger and older adults (e.g., Hasher and Zacks, 1979), whereas over-additive effects suggest that age-related deficits in terms of response speed interacts with semantic activation in order to produce greater savings in time when both the prime and target are semantically related (Laver and Burke, 1993).

But this does not mean that raw RT is always the most appropriate dependent variable. Other theoretical positions assume a different relationship between RT and mental operations that is most appropriately measured by a transformation such as log or inverse RT. For example, differences calculated on the logarithmic metric reflect proportional change [i.e., log(700 ms)−log(600 ms) = log(700/600 ms)], which aligns with many theories of aging which attribute a causal role to general cognitive slowing (e.g., Salthouse, 1985). However, the vast majority of cognitive theories have been developed and validated on raw RT. So by routinely applying a transformation to yield the normal distribution required for LMM, the researcher may ultimately fail to test their hypotheses using the dependent variable that underpinned their theoretical predictions.

In individual differences research, scale dependent interactions touch upon even broader theoretical implications. At its most basic conceptualization in a two-factor design, a significant interaction indicates that the effect of a particular variable (the numerical difference on the dependent variable between levels of one of the factors) changes across the population of interest because it differs as a function of a second independent variable; typically another group of people or a different condition. Conversely, a lack of interaction between these factors suggests that the average effect remains uniform across individuals or conditions under assessment. Thus, statistical assessment of interactions provides insight as to whether there is a single "true value that we are trying to approximate when we measure humans on some dimension" (Speelman and McGann, 2013, p. 2), or whether multiple values exist particular to each individual.

Thus, the increasing reliance on LMM in cognitive psychology presents researchers with a conundrum created by the mismatch between the dependent variable dictated by theory and the dependent variable dictated by the requirements of the statistical analysis. As discussed above, in cognitive psychological investigations of "mental chronometry," raw untransformed RTs are usually the metric about which the researcher has predictions. However, to satisfy the assumptions of LMM, the statistical analysis is conducted on the transformed metric. Thus, in order to interpret the results and in order to compare them with earlier published ANOVA data, the estimates of the empirical effects from the LMM are often "back-transformed" into raw RT. But unfortunately, back-transformation can be unreliable because statistically significant differences on the transformed metric are uninformative as to whether significant differences exist on the original untransformed metric and vice versa (Berry et al., 2010). Cognitive psychologists are therefore trapped between a rock and a hard place. Analyses on raw RT are inappropriate because they fail to meet the assumptions of the linear model, but analyses on transformed RT are uninformative because they fail to answer the research questions of interest.

The ideal solution to this quandary would be to allow statistical assessment on the original raw RT metric, but to also meet the mathematical constraints imposed by the statistical model. Such a solution is offered by generalized linear mixed-effect models (GLMMs) which offer one approach to achieving this ideal that is readily implemented in many statistical packages. By separating the mathematical and theoretical components of the model, GLMMs allow researchers to use the dependent variable most appropriate to their research question, while simultaneously meeting the mathematical criterion of normalized, homoscedastic residuals in linear regression. To achieve these goals, GLMMs require the researcher to consider these issues as part of the specification process.

A Case Study: Effects of Word Frequency and Stimulus Quality on Lexical Retrieval

To demonstrate the interpretative problems associated with routinely transforming RT to meet the normality assumptions of LMM and to illustrate how GLMM can be applied to avoid the need for transformation, we present re-analyses of data recently reported by Balota et al. (2013). Specifically, they used LMM to re-analyse the data from three published studies which reported additive effects of word frequency and stimulus quality in ANOVA analyses of raw RT (Yap and Balota, 2007; Yap et al., 2008). However, for the LMM analyses on inverse RT, the data transformation that most effectively normalized the residuals for all datasets, the results yielded a completely different pattern for all three experiments: significant underadditive interactions.

In "chronometric" research, additive or interactive effects reflect fundamental assumptions about the nature of RT described at the beginning of this paper. Because each measured

RT is assumed to reflect a composite of several distinct stages of processing, separate stages in mental operation can be inferred if the time required to perform a second mental operation is independent of the time required to complete the first mental operation (i.e., the effects are additive) (Sternberg, 1969). This reasoning is crucial for additive-factors logic (Sternberg, 1969), because without the ratio measurement scale properties in raw RT (Townsend, 1992), the inferential power of this technique is lost because equivalence in measurable raw RT can no longer be taken as evidence of equivalence in processing.

Thus, within the additive-factors logic (Sternberg, 1969) framework described above, the temporal relationship between word frequency and stimulus quality has important theoretical implications regarding the nature of lexical representation. Taken individually, low frequency words and visually degraded stimuli both serve to slow RT relative to when the stimuli are clearly presented or of high frequency (Stanners et al., 1975). However, the additive effects of these two variables on raw RT reported in the original papers suggest that that these factors selectively influence separate stages of mental processing, and produce significant challenges for activation models which predict interactive effects between frequency and stimulus quality (Borowsky and Besner, 1993). Specifically, activation models propose that the threshold for activation is determined by word frequency and the rate of activation by stimulus quality, so stronger effects of stimulus quality on low frequency words should therefore be observed because more time is required to reach the higher activation threshold for low frequency words when combined with a slower rate of activation in the context of degraded stimuli (Morton, 1969). This consistent evidence of additive effects of word frequency and stimulus quality in the experimental data, under conditions that yield interactions between each of these variables and semantic priming, therefore presents a strong challenge to fully interactive activation models (Borowsky and Besner, 1993; Balota et al., 2013). Given the central theoretical importance of the additive effects of word frequency and stimulus quality observed on raw RT, Balota et al.'s (2013) demonstration that the additive pattern is specific to raw RT and changes when the dependent variable is transformed directly reflects the theoretical quandary presented above.

The Generalized Linear Mixed-Effect Model (GLMM) Framework

GLMMs combine and extend the properties of LMM and generalized linear model (GLM) approaches, by relaxing LMM's assumption that the dependent variable (and the residuals) follow a normal (Gaussian) distribution, and extending GLM's scope of inference to extend beyond a single random population. Rather than making the default assumptions of LMM methods, GLMM requires researchers to specify a number of components of their data and design:

(1) the explanatory variables responsible for systematic variation in responses: referred to as the *fixed factors*;
(2) the sampling structure of the design contributing to random variability in responses: the *random factors*;
(3) the probability distribution describing the plausible processes underlying the observed data: the distribution of the *dependent variable*; and
(4) the mathematical function characterizing the relationship between the fixed factors and the dependent variable: the *link function*.

The following sections introduce the key theoretical and methodological issues regarding specification of GLMMs within the context of the three experiments from Balota et al. (2013). Readers interested in more technical mathematical and computational details regarding LMM (Pinheiro and Bates, 2000; Raudenbush and Bryk, 2002; Baayen, 2008), GLM (McCullagh and Nelder, 1989), and GLMM (Jiang, 2007; Stroup, 2013) should consult the excellent resources already published on these topics.

The three experiments re-analyzed by Balota et al. (2013) each factorially manipulated word frequency and stimulus quality within a lexical decision task. For the word responses in all three experiments, each participant responded to 100 high frequency and 100 low frequency words, presented in either clear or degraded stimulus quality conditions. In Yap and Balota (2007), the stimulus quality manipulation was conducted between subjects while Yap et al. (2008, Experiments 1 and 2) used within-subjects manipulations conducted on counterbalanced item sets. The non-word items in Yap and Balota (2007) and Yap et al. (2008, Experiment 1) comprised of 200 pronounceable pseudo-words (e.g., *flirp*), while Yap et al. (2008, Experiment 2) used 200 pseudo-homophones (e.g., *brane*). Further details regarding the design are available in each experiment's respective published reports.

The Fixed Factors

Users of ANOVA and ordinary least squares regression in the basic linear model framework will already be familiar with specifying fixed factors in their analyses. Both at a conceptual and practical level, this remains unchanged in GLMM. In order to differentiate them from random factors described below, fixed factors are the components of the linear predictor responsible for systematic variability in the observed responses. Typically, fixed factors consist of the independent variables (or covariates) with a small finite number of levels under experimental manipulation. The levels of these factors are the object of hypothesis testing (fixed effects), and represent the conditions for which the model provides estimates of the average response over the entire population(s) (generally denoted by the symbol $\widehat{\mu}$ —the estimated mean corresponding to each condition).

Across the three experiments reported in Balota et al. (2013), the fixed factors correspond to word frequency and stimulus quality. Normalized sum contrasts specified on these fixed factors yielded four fixed effects in the statistical model: mean RT associated with the lexical decision task (intercept), differences in RT associated with the manipulations of word frequency (high vs. low), stimulus quality (clear vs. degraded), and frequency ×

stimulus quality interaction[1]. Of central interest is whether the observed data are consistent with interactive effects of frequency and stimulus quality predicted by interactive models, or the additive effects that follow from the independent processing stages assumed by serial models.

The Random Factors

Within the mixed modeling framework, random factors correspond to components of the linear predictor in which a random subset of levels are sampled from a larger population. As opposed to fixed factors, in which systematic variability between conditions (i.e., mean differences) is explicitly estimated and compared, variability in the random factors is used to: (1) estimate the extent to which mean responses vary across units of the random factor; (2) allow inferences about whether fixed effects generalize beyond the units sampled in the random factor; (3) remove variability in responses that are associated with the random factor rather than the conditions of experimental interest (i.e., reduce Type I error rate). Typically, many levels of the random factor are sampled in the experiment under which responses are meaningfully clustered. Although clustering is one form of structural dependency typically associated with a random factor, other structural dependencies such as nesting, cross-classification, blocking and other counterbalancing procedures can also contribute to nuisance variability that is partialled out with a random factor[2].

Subjects and items constitute the random factors common across the three experiments reported in Balota et al. (2013), because responses are clustered according to individual participants and English words—both of which represent a random sample from their respective populations. Following nomenclature within the LMM literature (e.g., Barr et al., 2013), the overall mean for each subject and item were estimated as "random intercepts" in each of the experiments, while with the degree to which each fixed effect varied across subjects and/or items were estimated as "random slopes." This latter specification for random slopes differed according to the design of the three experiments. In the Yap and Balota (2007) experiment, stimulus quality was manipulated between-subjects and word frequency was manipulated between-items, so the random slopes controlled for subject-specific variability in the frequency effect which can be distinguished from variability associated with particular words, and item-specific variability in the stimulus quality effect which can be distinguished from variability associated with different participants. For the other two experiments in which word frequency and stimulus quality were both manipulated within-subjects, the random slopes controlled for subject-specific variability in the frequency effect, stimulus quality effect, and frequency by stimulus quality effect, as well as item-specific variability in the stimulus quality effect. This represents the "maximal" random effect structure (Barr et al., 2013) for each of the experiments.

The Dependent Variable

A key feature of GLM and GLMM is the ability to appropriately model a variety of response distributions. As noted previously, GLMM does not make the default assumption that this distribution is Gaussian and therefore requires that the researcher specify an appropriate distribution. In some measurement contexts, this selection is straightforward—binary responses are described by a binomial distribution; count responses are described by a Poisson distribution. But selecting the appropriate dependent variable is less straightforward in domains like cognitive psychology, where researchers often investigate latent constructs that are indexed by continuous behavioral measures, like RT, which can be described by a host of distributions (e.g., normal, beta, gamma, uniform, etc.), and where there is often no consensus on the "correct" distribution. This ambiguity has contributed to researchers' willingness to transform RT measures to meet the mathematical assumptions of LMM. GLMM offers an alternative: the researcher can select the quantitative distribution that best captures the properties of their measured variable. As we describe below, both theoretical and empirical considerations underpin this decision.

Across the three experiments reported in Balota et al. (2013), the dependent variable was the RT to correctly classify each stimulus as an English word. As illustrated in **Figure 1**, the distributions of observed RT (represented by solid lines) for all three experiments were unimodal with a distinct positive skew. In addition to this characteristic shape, the data for all experiments also revealed a linear relationship between the standard deviation of RTs and mean RT demonstrated in many previous studies of RT in binary choice tasks (e.g., Luce, 1986; Faust et al., 1999; Wagenmakers and Brown, 2007). This linear relationship is also evident in plots of the residuals which show hetereoscedasticity in LMM analyses, evidenced by increasing spread in residuals for longer predicted RT (Kliegl et al., 2010; top row of plots in **Figure 3**).

Rather than transforming the dependent variable to eliminate this deviation from normality, GLMM allows the researcher to select a theoretical distribution that matches the properties of measured RT. Two of the two-parameter distributions currently implemented for GLMMs in the stats package as part of the default installation of the R program for statistical computing (R Core Team, 2013), the Gamma and Inverse Gaussian distributions reproduce these surface characteristics of raw RT—a unimodal skewed distribution with continuous responses greater than or equal to 0. As shown in **Figure 1**, they both provide a closer approximation to the observed distribution of RTs

[1] Balota et al. (2013) also included the lexicality and stimulus quality of the previous trial as fixed factors in their analyses in order to investigate the modulating role of trial history on performance, and to assess the generality of Masson and Kliegl's (2013) claim that additive effects of word frequency and stimulus quality are a spurious outcome of ignoring trial history. Evaluating the effects of such trial level variables is only possible in LMM and GLMM using unaggregated data because they allow structural dependencies to be accounted for as random factors. However, Balota et al. (2013) reported no evidence of previous trial history significantly modulating the relationship between word frequency and stimulus quality, so these variables were not included in our analyses.

[2] At the time of writing, implementation of LMM and GLMM in popular statistical software assumes that the mean responses across the units of the random factor are normally distributed. Though this may be a reasonable assumption given that sample means can be normally distributed even though the underlying population of responses is non-normal based on the central limit theorem, further advances in computation may allow non-normally distributed random factors to be specified in doubly generalized linear mixed-effect models as described by Lee et al. (2006).

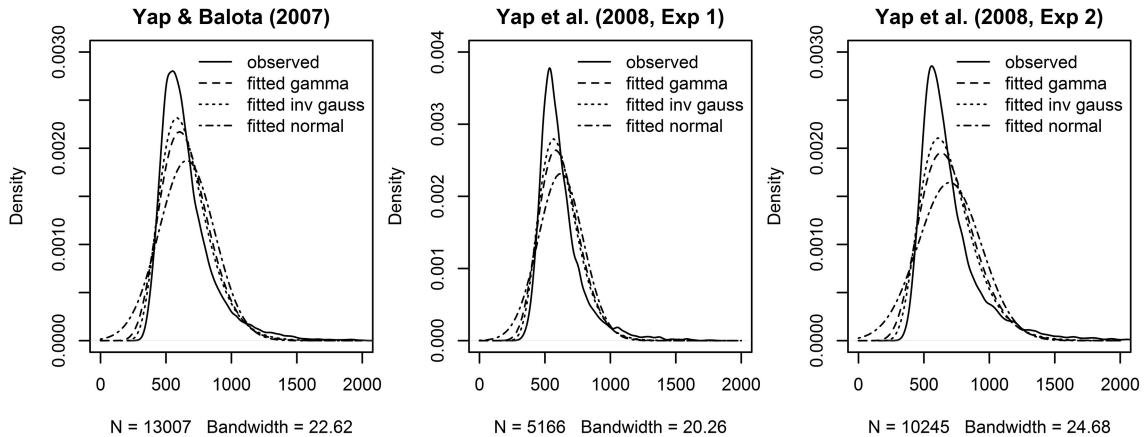


FIGURE 1 | Observed (solid lines) and fitted densities (dotted, dashed and broken lines) to the RT from the three Balota et al. (2013) experiments. For the Yap and Balota (2007) experiment, the gamma density was fitted with a rate of 0.018 and a shape parameter of 11.967 (dashed line), the inverse Gaussian density was fitted with a λ parameter of 8201.5 and a mean of 658.8 (dotted line), the normal density was fitted with a mean of 658.8, and a standard deviation of 212.97 (broken line). For the Yap et al. (2008, Experiment 1), the gamma density was fitted with a rate of 0.026 and a shape parameter of 16.000 (dashed line), the inverse Gaussian density was fitted with a λ parameter of 10223.8 and a mean of 620.1 (dotted line), the normal density was fitted with a mean of 620.1 and a standard deviation of 172.26 (broken line). For the Yap et al. (2008, Experiment 2), the gamma density was fitted with a rate of 0.015 and a shape parameter of 10.740 (dashed line), the inverse Gaussian density was fitted with a λ parameter of 7704.3 and a mean of 697.0 (dotted line), the normal density was fitted with a mean of 697.0 and a standard deviation of 242.67 (broken line).

in the three experiments than the normal distribution. The distributions also provide an explicit mathematical relationship between the mean and variance. For the Gamma distribution, the variance of the distribution increases proportionally with the mean, while the variance increases proportionally with the cube of the mean for the Inverse Gaussian distribution. Despite the differences in their mathematical expression, both distributions are able to approximate a variety of distributional shapes that allow them to "statistically mimic" RT responses and yield fits that are practically indistinguishable from each other (Van Zandt and Ratcliff, 1995).

As well as approximating the surface characteristics of the distribution of the dependent variable, the probability distribution should also provide a plausible description of the processes underlying the response. At a conceptual level, both the Gamma and Inverse Gaussian distributions can be linked to *waiting time*—how long it takes until an event of interest (e.g., a button press) to occur. Mathematically, the Gamma distribution is the sum of multiple exponential distributions, which can be considered to model the probability that no event occurs until a certain period of time. The Gamma distribution can therefore be considered to model several serial stages of processing, each of which finishes with a time that is exponentially distributed (Van Zandt and Ratcliff, 1995). Similarly, the Inverse Gaussian distribution has been identified with the time for evidence accumulation to reach a single threshold boundary within a diffusion process (Schwarz, 2001).There are other distributions as described in the General Discussion (e.g., ex-Gaussian, ex-Wald, shifted Wald) with parameters that have also been associated with psychological processes underlying RT (Matzke and Wagenmakers, 2009). Given that there is no consensus as to the "correct" distribution for mapping from psychological processes to RTs, the purpose of this introduction is not to advocate for a particular distribution, but to illustrate that the Gamma and Inverse Gaussian are examples of distributions that provide a plausible description of processes reflected in RT.

The Link Function

In GLM and GLMM, fixed factors are assumed to be linear predictors of a function of the observed response rather than the observed response itself. Thus, the model assesses the linear predictors ($\widehat{\mu}$) on an *unbounded transformed scale* (e.g., the scale upon which a latent variable like "lexical retrieval" is measured could contain any numerical value), that is different from the *bounded original scale* of the dependent variable (DV) (e.g., observed RT contains strictly positive values like those produced by the Gamma distribution; the observable probability of an inaccurate response is bound between the values of 0 and 1 like those from a binomial distribution). The transformed and original scales are connected by a monotonic differentiable link function that allows back-transformation to the original metric by providing a one-to-one mapping between the range of fitted values produced by the linear predictor on the transformed metric and the range of observed values on the original metric [i.e., DV = $f(\widehat{\mu})$]. Therefore, the nature of the relationship between the two scales can be considered to be defined by the mathematical function connecting the observed response to the latent construct upon which the fixed factors are assessed. In the special case where "no function" is required and the observed response is assumed to directly tap the latent construct (e.g., RT is a direct measure of the time required for lexical retrieval), the function binding the expected values produced by the predictors to the dependent variable is the *identity link* (i.e., DV = $\widehat{\mu}$). Ordinary linear regression and LMM assumes an identity link

between the DV and the latent construct. When researchers using these methods believe that the measured DV is not directly related to the latent construct, they can mathematically transform the DV into the latent construct, and then apply this transformed variable in the analysis as the DV in order to achieve a similar effect[3]. That is, the link function in GLM(M) explicitly defines the nature of the expected relationship between the predictors and the observed response.

In the context of the experiments reported in Balota et al. (2013), there are two reasons as to why the identity link is appropriate. Firstly, from a theoretical perspective, the tradition of mental chronometry assumes that manipulations directly affect RT rather than some function of RT. More explicitly within additive factors logic, RT is assumed to be linearly affected by the experimental factors so that factors that affect a single processing stage interact, while those that affect separate processing stages do not. By changing the form of this mapping with a non-linear link function or a non-linear transformation of the dependent variable as applied in LMM, such an interpretation cannot be applied and cannot inform the models from which they were derived. Secondly, from a mathematical perspective, a non-linear link function is usually applied to constrain the predicted values within the bounds of the dependent variable. Since the bulk of observed RTs in Balota et al. (2013) are situated well away from the negative boundary (in part because RTs faster than 200 ms were removed), and predictions are not extrapolated beyond the observed conditions, there is little danger of the model producing impossible negative values for RT which are difficult to interpret.

Using GLMM to Avoid the Need for Transformation of Skewed RT Data

To illustrate the application of GLMM to address the problems with transformation outlined earlier, we re-analyzed the three experiments that Balota et al. (2013) recently demonstrated to yield contradictory outcomes in analyses conducted on raw and transformed data. They report that LMM analyses of the inverse RT transformed data that best satisfied criteria for normality yielded underadditive interactions rather than the additive effects of frequency and stimulus quality found with raw RT.

We report the results of six analyses of each of the three experiments. Two of the analyses parallel Balota et al.'s (2013), by using LMMs on raw RT (DV = RT) and inverse RT (DV = -1000/RT). By default, these analyses assume a Gaussian distribution and identity link function. The remaining four analyses are GLMMs on raw RT which assume either a Gamma or Inverse Gaussian distribution of the DV, and a linear (identity link function; RT = $\widehat{\mu}$) or inverse relationship (inverse link function; RT = $-1000/\widehat{\mu}$) between the predictors and RT. We chose $-1000/\widehat{\mu}$ as the specific form of the inverse link function to parallel the inverse transformation applied to RTs in Balota et al.'s (2013) LMM analyses (i.e., -1000/RT). These GLMM analyses are therefore analogous to the LMM analyses conducted on inverse RT.

The primary interest is in the results from the properly specified GLMM based on the decisions described in the previous section, but we also aim to clarify how differences in the specification of the dependent variable and link function relate to the conflicting findings between raw and inverse transformed RT reported by Balota et al. (2013).

The analyses were conducted on RT data for correct word responses for Yap and Balota (2007) and Yap et al. (2008 Experiments 1 and 2) using version 1.0-5 of the lme4 package (Bates et al., 2013) in the R program for statistical computing (R Core Team, 2013) following the same trimming procedures described in Balota et al. (2013). Since there is continuing debate as to how *p*-values should be generated for LMMs because of computational issues regarding degrees of freedom, we follow the current practice of considering effects greater than two standard errors (i.e., $|t|> 2$) to be significant at the 0.05 level for datasets involving a large number of observations (Kliegl et al., 2010; Masson and Kliegl, 2013). The R syntax used to generate these models along with the full model output and predicted mean RT for each condition can be found in the Supplementary Materials.

Figure 2 summarizes the predictions of the models assuming a linear relationship between the predictors and RT for the three experiments. The corresponding results for models assuming an inverse relationship between the predictors and RT are presented in **Figure 4**. Each column of **Figures 2–5** corresponds to a different experiment, while the rows of the figures present estimates from the LMM models (top row), and GLMM models assuming Gamma (middle row), and Inverse Gaussian (bottom row) distributions, respectively, of the DV.

For each model summarized in **Figures 2**, **4**, the shaded region of the prediction plot depicts the estimated effect of word frequency (difference between high and low frequency conditions) based on the fitted values for each of the four frequency by stimulus quality conditions as plotted on the model transformed scale (x-axis), while the y-axis plots the same difference after the mean estimates have been back-transformed via the link function on the original RT scale. The estimates are identical on the model and back-transformed RT scales in **Figure 2** because the identity link assumes that the scale of the latent construct assessed by the model (x-axis) is synonymous with RT. The form of the link function itself is depicted by the solid black line connecting the diagonals of the plot.

Although an identity link function (DV = $\widehat{\mu}$) was also specified for the LMM analysis on inverse transformed RTs (DV = -1000/RT), we depict a non-linear function in **Figure 4** to illustrate the back-transformation from inverse to raw RT (RT = $-1000/\widehat{\mu}$) that researchers typically apply to interpret their data. The *p*-value corresponding to the critical interaction effect, which is presented in the bottom-right corner of each plot only assesses whether there is a significant difference in the linear effect of frequency on the model transformed scale (x-axis), and does not assess whether significant (linear) differences exist on the original RT scale (y-axis) unless the identity link was specified (Berry et al., 2010).

[3]It is important to note that differences in the logs of the means (i.e., passing $\widehat{\mu}$ through a log link) is not the same as differences in the means of log-transformed data, but general compression in differences involving larger values on either scale is maintained in either method.

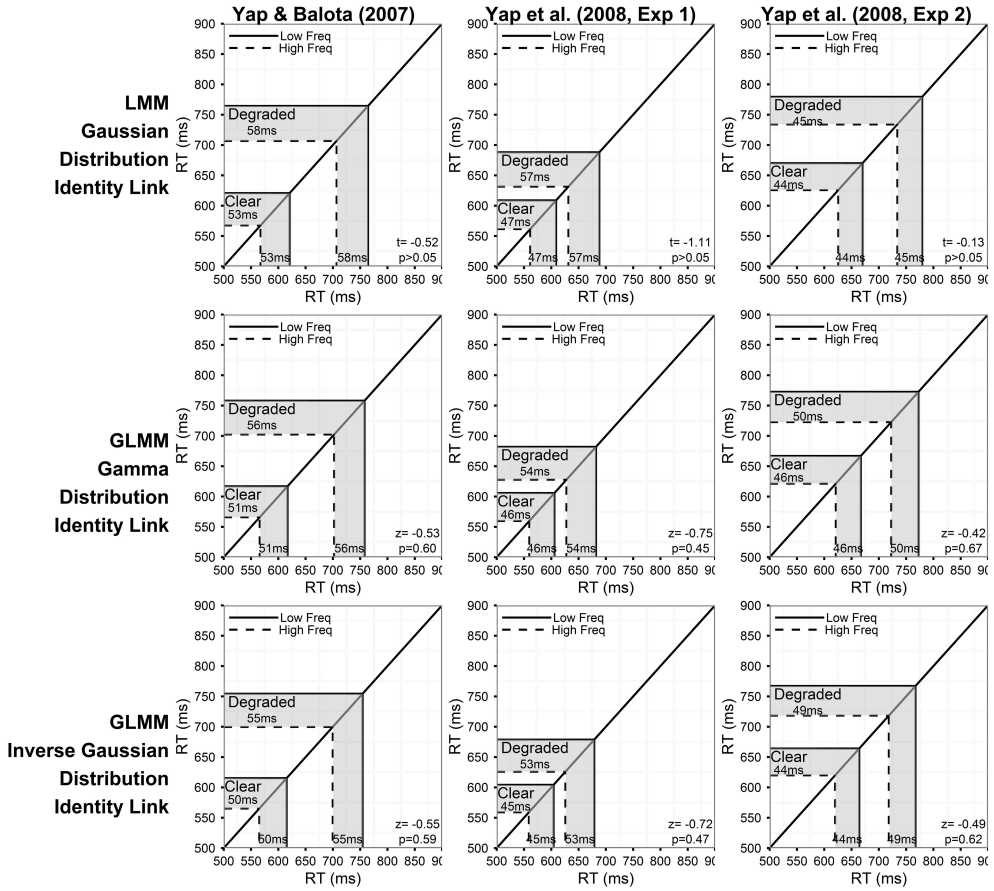


FIGURE 2 | Prediction plots illustrating the estimated frequency effect and statistical results (*t*- or *z*- value and corresponding *p*-value) of the word frequency by stimulus quality interaction (shaded region on x-axis) based on models assuming a linear relationship between the predictors and RT (identity link function). Note that the back-transformed estimates (shaded region on y-axis) are identical because of the identity link function. Each column of plots represents the results from a different experiment (from left to right: Yap and Balota, 2007; Yap et al., 2008: Experiment 1; and Yap et al., 2008: Experiment 2), and each row of plots represents a different assumption for the distribution of RTs (from top to bottom: Gaussian, Gamma, and Inverse Gaussian). Note that precise *p*-values are produced in GLMM for the Wald *Z*-statistic in R, while approximate *p*-values can only be inferred based on the magnitude of the *t*-value produced in LMM.

Selecting the Right Model

Each of the individual analyses in **Figures 2**, **4** produced subtle differences in the magnitude, direction or statistical significance of the word frequency by stimulus quality interaction. A decision must therefore be made about the best-fitting correctly specified model. There are a number of ways to address this question.

Throughout the previous sections, we have argued that, from a theoretical perspective, the dependent variable of theoretical interest in mental chronometric research like this is raw RT, and that additive factors logic assumes a linear relationship between the experimentally manipulated variables and RT itself. From this perspective, only the analyses using raw RT as the dependent variable and specifying an identity link function provide meaningfully interpretable results for this experiment (**Figure 2**).

To further discriminate between the analyses, we can identify the statistical model that provides predictions which best fits the observed RTs. **Figure 3** allows a visual inspection of model fit, by plotting the residuals against predicted RT. The LMM analyses (top row of plots), which assume a Gaussian distribution of raw RT, clearly exhibit a heteroscedastic (fan-shaped) pattern that is not evident in the GLMM analyses assuming a Gamma or Inverse Gaussian distribution (middle and bottom row of plots). Therefore, these plots suggest that the Gamma or Inverse Gaussian distributions provide a better fit to the data because they explicitly account for the hetereoscedastic pattern of increasing variability with slower responses and therefore yield more normally distributed residuals.

A similar conclusion derives from AIC and BIC summary fit indices presented in **Table 1**, and the estimated Gaussian, Gamma, and Inverse Gaussian distribution fits to the observed RT density in **Figure 1**. Across the three experiments, the Inverse Gaussian distribution (followed by the Gamma and Gaussian distributions) produce parameters that best approximate the shape of the observed RT distribution, and yield fit values

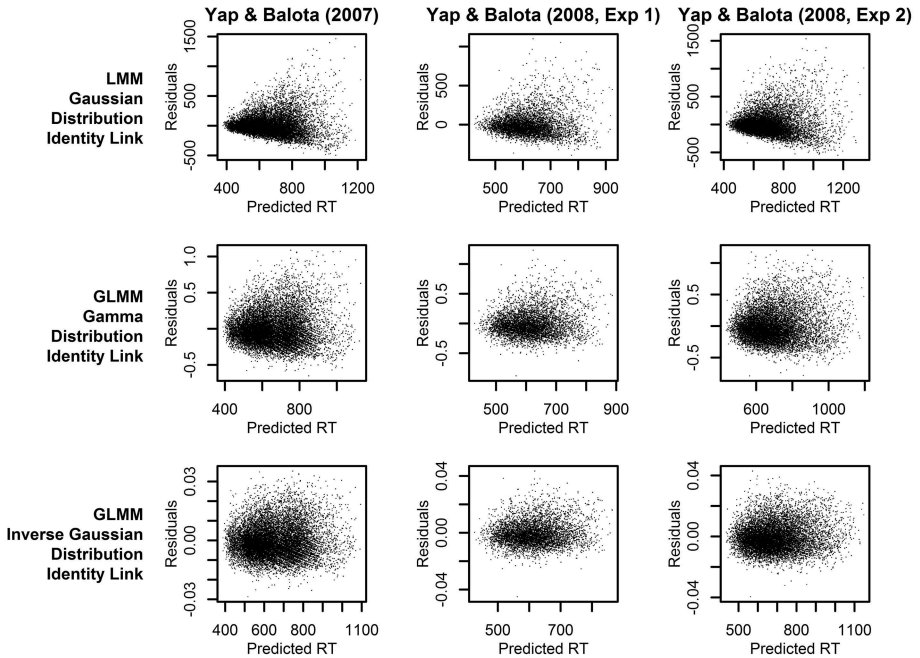


FIGURE 3 | Plots of the residuals over predicted RT from models assuming a linear relationship between the predictors and RT (identity link function). Each column of plots represents the results from a different experiment (from left to right: Yap and Balota, 2007; Yap et al., 2008: Experiment 1; and Yap et al., 2008: Experiment 2), and each row of plots represents a different assumption for the distribution of RTs (from top to bottom: Gaussian, Gamma, and Inverse Gaussian).

that are consistently lower than the Gamma or Gaussian distributions. Thus, on both these graphical and empirical indices, the Inverse Gaussian distribution provides the best fitting model.

Having identified the most appropriate statistical model, we can consider its results. Consistent with the ANOVA analyses reported in the original published papers, none of the three experiments yielded a significant interaction between word frequency and stimulus quality in the Inverse Gaussian GLMM with identity link function (bottom row of plots in **Figure 2**). This model predicted effects of frequency that were 5, 8, and 5 ms greater for the degraded than clear condition in the Yap and Balota (2007), Yap et al. (2008, Experiment 1), and Yap et al. (2008, Experiment 2) data, respectively. The magnitude and direction of these effects are essentially identical to the 6, 7, and 5 ms overadditive effect reported in original ANOVA analyses. Although these estimated effects are similar to those predicted in the poorer fitting Gamma and Gaussian GLMM with identity link (top and middle row of plots in **Figure 2**), the test statistic (*t*- or *z*-value) is larger and corresponding *p*-value lower for the better fitting models, suggesting that the standard errors have been more precisely estimated. Better fitting models provide more powerful adjustment to extreme values, particularly in the slowest condition of degraded low frequency words, where calculation of the average would be most affected, thus allowing greater power as well as reliability with which to assess individual differences between subjects and items (see Appendix in Supplementary Material for mean RT predicted for each condition by the six models).

Different conclusions about the relationship between word frequency and stimulus quality are suggested by the results of models using transformed RTs or link functions that assume a non-linear relationship between the predictors and RT. From the perspective of model fit alone, the analysis on inverse transformed RT produces residuals that offer the least amount of hetereoscedasticity (**Figure 5**), suggesting that the fit is at least as good, if not better, than the Inverse Gaussian GLMM with identity link described above[4]. This is the expected outcome of applying the Box–Cox procedure to estimate a power transformation that stabilizes variance in order to create normally distributed data. However, although these models meet the mathematical assumptions of normality required by LMM, as Balota et al. (2013) report, relying on the transformed DV in LMM put the researcher in the unhappy situation of developing an *ad-hoc* explanation of why the estimated effect of frequency is now underadditive (**Figure 4**), as opposed to the additive or slightly overadditive effects observed on raw RT.

These contradictions arise because interval differences in the dependent variable are distorted when non-linear transformations are applied. For each of the prediction plots based on an inverse transformation or inverse link function in **Figure 4**, almost all of the back-transformed estimates suggest no difference, or a slightly larger numerical effect of frequency for degraded words (a small overadditive effect) on the RT scale (y-axis). However, on the model estimate scale

[4]Empirical fit indices such as AIC/BIC values are not comparable across models with different dependent variables (Burnham and Anderson, 2002).

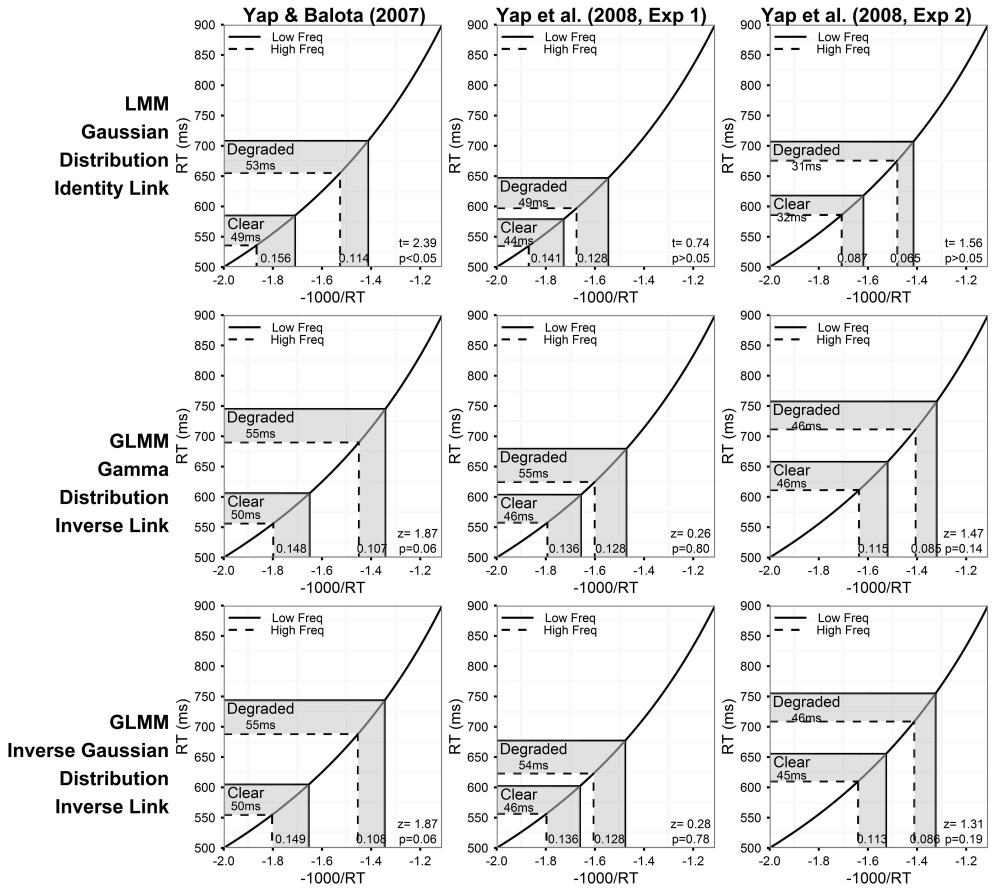


FIGURE 4 | Prediction plots illustrating the estimated frequency effect and statistical results (*t*- or *z*-value and corresponding *p*-value) of the word frequency by stimulus quality interaction (shaded region on x-axis) based on models assuming an inverse relationship between the predictors and RT (inverse link function). The plots also present the back-transformed estimates (shaded region on y-axis) on the original RT metric. Each column of plots represents the results from a different experiment (from left to right: Yap and Balota, 2007; Yap et al., 2008: Experiment 1; and Yap et al., 2008: Experiment 2), and each row of plots represents a different assumption for the distribution of RTs (from top to bottom: Gaussian, Gamma, and Inverse Gaussian). Note that precise *p*-values are produced in GLMM for the Wald Z-statistic in R, while approximate *p*-values can only be inferred based on the magnitude of the *t*-value produced in LMM.

(x-axis), these differences are distorted by the non-linear inverse link function into a numerically larger effect of frequency for clear words (underadditive effect). For the Yap and Balota (2007) experiment, the distortion caused by the non-linear transformation was severe enough to push the underadditive effect to statistical significance in the LMM analysis (top left panel of **Figure 4**). The underadditive interactions in this dataset were also marginally significant in the GLMM analyses using the inverse link function.

To meaningfully interpret this underadditive effect, and effects assessed on the inverse RT scale more generally, the researcher must assume that the predictors are inversely related to RT. This view is consistent with recent attempts to map effects assessed on the reciprocal scale to differences in processing rate or processing speed (Kliegl et al., 2010). For example, processing rate or speed of evidence accumulation is assumed to be slower for visually degraded as opposed to clearly presented words in activation models (e.g., McClelland and Rumelhart, 1981), thus yielding the slower RT typically observed for these conditions. However, a core assumption within all of these models is that rate of evidence accumulation is linear over time (e.g., Borowsky and Besner, 1993; Ratcliff and Rouder's, 2000, diffusion model; Brown and Heathcote's, 2008, linear ballistic accumulator)—in direct contrast to the non-linear relationship implied by the inverse scale. So while there may be physiological reasons to expect non-linearity at the level of neural spike rates (e.g., Carpenter and Williams, 1995), the implications associated with the reciprocal nature of this transformation on RT appears to be limited because psychological models assuming linearity are able to closely predict responses in observed data (Ratcliff, 1978; Brown and Heathcote, 2008).

Thus, the GLMM procedure allows researchers to select the DV most appropriate to their research question rather than use a transformed DV simply to meet mathematical assumptions. If raw RT is the most appropriate metric, as we have argued to be the case for most mental chronometric research, an

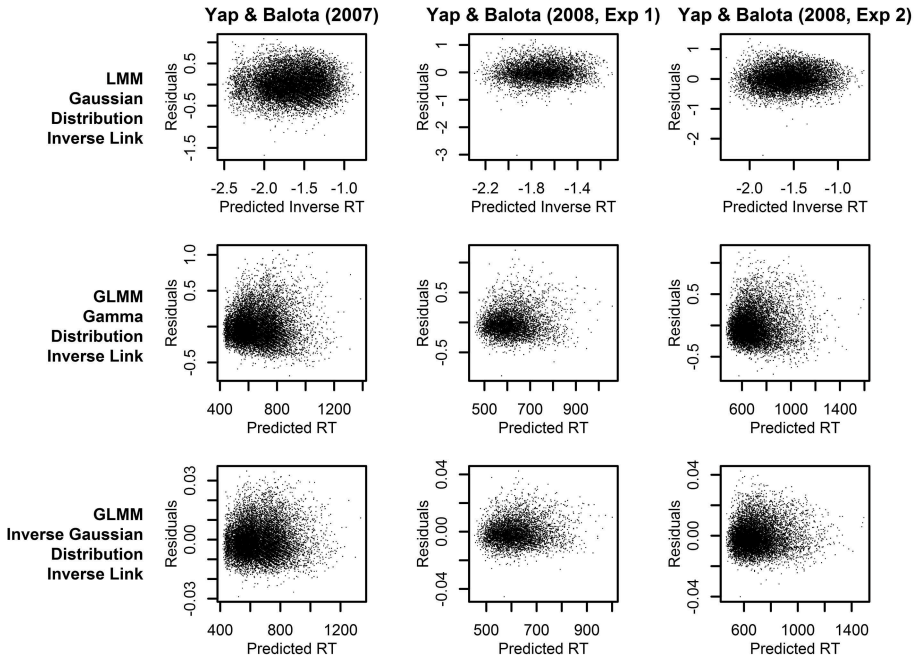


FIGURE 5 | Plots of the residuals over predicted RT (or inverse RT) from models assuming an inverse relationship between the predictors and RT (inverse link function). Each column of plots represents the results from a different experiment (from left to right: Yap and Balota, 2007; Yap et al., 2008: Experiment 1; and Yap et al., 2008: Experiment 2), and each row of plots represents a different assumption for the distribution of RTs (from top to bottom: Gaussian, Gamma, and Inverse Gaussian).

TABLE 1 | AIC and BIC indices of model fit comparing LMMs and GLMMs of different distribution and link assumptions for each of the three experiments.

Link function	Distribution (DV)	Yap and Balota (2007)		Yap et al. (2008, Experiment 1)		Yap et al. (2008, Experiment 2)	
		AIC	BIC	AIC	BIC	AIC	BIC
LMM (Identity link)	Gaussian (inverse RT)	6337	6404	3284	3356	6832	6912
	Gaussian (raw RT)	170,573	170,640	66,775	66,847	138,196	138,276
GLMM (Identity link)	Gamma (raw RT)	164,722	164,790	64,954	65,026	133,528	133,608
	Inverse Gaussian (raw RT)	163,161	163,229	64,461	64,533	132,318	132,398
GLMM (Inverse link)	Gamma (raw RT)	164,545	164,613	64,870	64,942	133,304	133,384
	Inverse Gaussian (raw RT)	163,012	163,079	64,395	64,467	132,128	132,207

Note that the dependent variable (DV) specified in the first row (LMM) were on inverse transformed RT, so these fit indices are not directly comparable with the other five rows of models which used raw RT as the DV.

Inverse Gaussian or Gamma distribution can be assumed to achieve more normal homoscedastic residuals, while retaining raw RT as the DV. As **Figure 2** shows, this produces more power than LMMs conducted on raw RT. Alternatively, if the researcher's predictions are for a transformed scale, such as inverse RT, specifying a non-linear link function of the same form as the inverse transformation applied to RTs (inverse link function; $-1000/\widehat{\mu}$) produces an identical distortion of frequency effects toward underadditivity (see middle and bottom row of prediction plots in **Figure 4**). Moreover, there appears to be no loss in model fit relative to the matching models using an identity link according to both a visual inspection of the residuals (**Figures 3**, **5**) and empirical fit statistics (**Table 1**).

In summary, GLMMs allow assumptions regarding the relationship between the predictors and the dependent variable to be tested independently of assumptions regarding the distribution of dependent variable. In LMM, the two are confounded because the relationship between the predictors and the dependent variable is dictated by the transformation selected to normalize the distribution of the dependent variable. By contrast, GLMM allows the form of the link function to be determined by the theoretical issues under consideration.

General Discussion

The broad goal of this paper is to echo Speelman and McGann's (2013) cautions about the routine use of statistical procedures without reflecting on the theoretical assumptions underlying their use. Within cognitive psychology, researchers are keenly aware of the dangers associated with relying on the mean, and many have begun to turn to the multilevel properties of LMMs as a way of simultaneously controlling for (or explicitly investigating) individual sensitivity between each item or participant as an explanation of overall differences between conditions (Clark, 1973; Locker et al., 2007). These methods offer one approach to reconciling the logic of group-based and individually focused research, one of the topics suggested for this Special Issue.

However, this change in statistical practice raises a new set of theoretical assumptions that have to be critically evaluated. Many cognitive researchers have adopted LMM because it is the statistical technique in current vogue, and a vast majority follow the recommendation to normalize RTs without proper consideration of the implications of such transformation for the theoretical rationale underpinning their research question. While for some researchers, the issues and recommendations proposed in this paper seem as obvious to those provided by Speelman and McGann (2013) with respect to the mean, we hope for many others that this discussion will serve as a timely reminder to reflect on the theoretical implications wedded to a seemingly innocuous statistical procedure.

Specifically, we have argued that raw RT is the most appropriate metric from the assumptions derived as part of the "mental chronometry" approach. However, transforming the dependent variable might be more appropriate from other theoretical perspectives. For example in the aging literature, theories of general cognitive slowing (e.g., Salthouse, 1985) propose that larger differences in RT for older as opposed to younger adults arise simply because the older adult's slower responses allow more time for the experimental effect to manifest (e.g., Kliegl et al., 2010). Such models therefore predict that the magnitude of effect expressed by younger and older adults should be defined by a constant ratio across RT (Myerson et al., 1992). Returning to the semantic priming example presented in the introduction, we showed that proportional differences can be mathematically expressed through logarithms. Thus, at a conceptual level, log RT is more appropriate than raw RT if one's research question is concerned with whether an experimental effect deviates from the theoretically defined proportional increase expected for slower responses. In our semantic priming example, parallel analyses of log and raw RT would therefore provide useful complementary insight regarding the nature of the relationship between response speed, age, and lexical activation.

There are, however, two major obstacles which impede the widespread application of logarithmic transformations within psychological data. The first is the finding in large-scale meta-analyses that proportional effects predicted by models such as general cognitive slowing are not fully captured by a logarithmic transformation alone, (e.g., Chapman et al., 1994; Faust et al., 1999). This is echoed in applications of the Box–Cox procedure in LMM analyses which typically identify the reciprocal rather than natural logarithm as the transformation best suited for psycholinguistic data (Balota et al., 2013). The result is that comparisons using LMM are being conducted on the inverse scale rather than on log or raw RT for which the researcher has predictions. By separating the mathematical issues related to the distribution of RT in GLMM, the researcher is able to specify the form of the link function (e.g., log, identity) that directly addresses their theoretical questions of interest.

The other major goal of the present paper is to introduce how GLMMs might be specified using a popular statistical program and concrete psycholinguistic example (see Appendix in Supplementary Material). Using a GLMM that fulfilled the mathematical requirements of homoscedastic residuals by assuming an Inverse Gaussian distribution but maintained the theoretically relevant dependent variable through the identity link function, the results yielded additive effects of word frequency and stimulus quality across the three experiments from Balota et al. (2013). This finding is important for two reasons. Computationally, the more powerful GLMM analyses yield statistical outcomes that confirm the robust additivity reported between these factors in previous literature, and yield numerical results that are consistent with a small overadditive effect estimated in the ANOVA analyses conducted by Yap and Balota (2007) and Yap et al. (2008). Theoretically, additive effects are consistent with separate stages of processing within the additive-factors framework (Sternberg, 1969) and support interpretations that assume an initial perceptual normalization process that is sensitive to stimulus quality which precedes the memory retrieval process responsible for effects of frequency (Borowsky and Besner, 1993; Yap and Balota, 2007).

Alternatively, additive effects of word frequency and stimulus quality can be accommodated in dynamic connectionist models (e.g., Plaut and Booth, 2000). A core assumption underlying these models is that the amount of activation required for the network to settle and output a RT response depends on the strength of its input along a non-linear sigmoidal function (see **Figure 6**). Variables which produce stronger input (e.g., higher frequency words, more semantically related concepts, older individuals with greater reading or perceptual ability) elicit stronger activation within the network, and thus output faster RT. However, proportionally smaller differences on RT are expected if all of the input falls within the upper and lower extremities of the sigmoid for which RT is most compressed (right part of **Figure 6**), relative to the more linear middle portion of the activation curve (left part of **Figure 6**). As described above, this proportional difference can be mathematically defined through a non-linear transformation. For example, a reciprocal relationship between input and RT (i.e., $RT = -1000/\hat{\mu}$ as in **Figure 4**) might characterize a situation in which the input strength associated with word frequency and stimulus quality are both assumed to fall at specific points within the lower rising part of the sigmoid. But in order to yield the observed additive effect on RT, a smaller effect of frequency must have arisen among the clearly presented items, which are assumed to produce stronger input. Given the positive relationship between input and

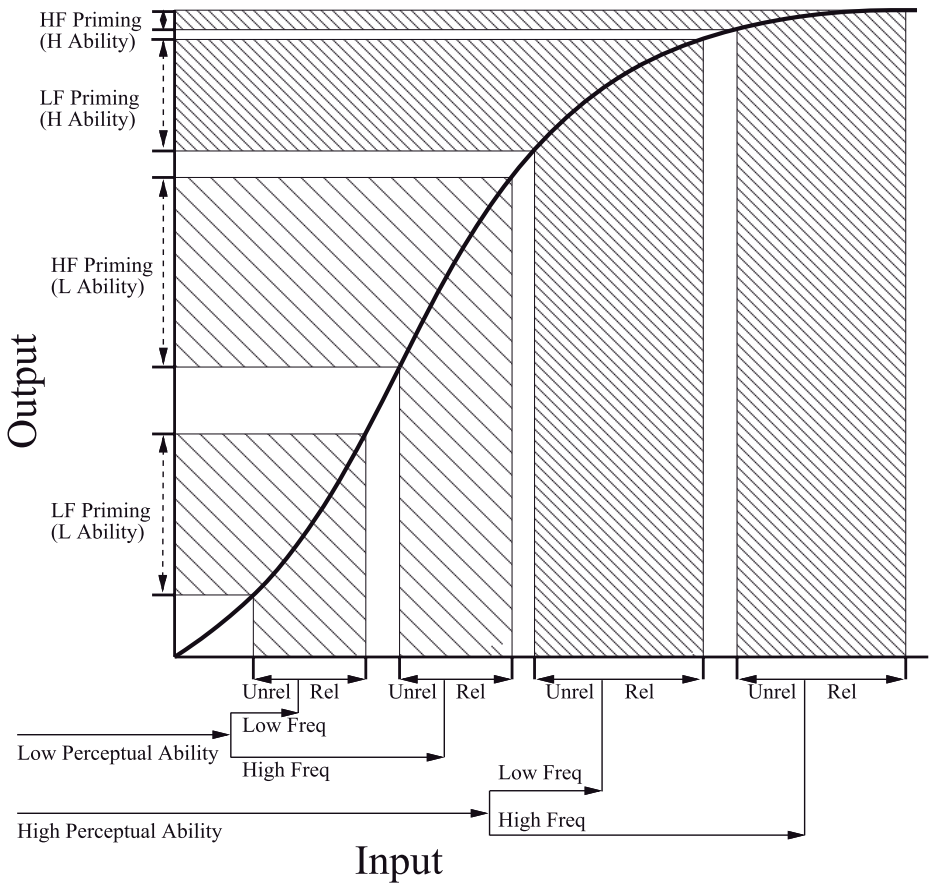


FIGURE 6 | The sigmoid activation function from the Plaut and Booth (2000) model. In this diagram the effects of perceptual ability, word frequency, and semantic priming are all assumed to lie in the linear and upper portion of the curve. From "Individual and developmental differences in semantic priming: Empirical and computational support for a single-mechanism account of lexical processing" by Plaut and Booth (2000). Copyright 2000 by American Psychological Association. Reproduced with permission.

activation, this finding is exactly opposite to that predicted by activation models as described in the Introduction.

Conversely, a completely opposite pattern is derived if the effects of word frequency and stimulus quality are both assumed to fall on the upper part of the sigmoidal function (as depicted in **Figure 6**). For example, specifying a logarithmic link function [RT $=$ 500 $\times$ log($\widehat{\mu}$)], paralleling the upper section of the sigmoid function within GLMM analyses assuming an Inverse Gaussian distribution of RT, revealed a trend toward significant overadditive interaction in all three experiments ($z = -1.75$, $p = 0.08$, for Yap and Balota, 2007; $z = -1.26$, $p = 0.21$, for Yap et al., 2008 Experiment 1; $z = -1.45$, $p = 0.15$; for Yap et al., 2008 Experiment 2). Individuals can therefore yield underadditive, additive or overadditive effects depending on their hypothesized position on the sigmoidal function.

As a concrete demonstration of this possibility, Plaut and Booth (2000) hypothesized that children of both high and low perceptual ability lie within the more linear portion of the sigmoid, because these less proficient readers are understood to possess generally weaker input than highly proficient adult readers. The result is that the magnitude of semantic priming is approximately equal for both high and low frequency words among those of high or low perceptual ability. In contrast, adult readers are hypothesized to possess greater input strength, positioning them within the upper part of the sigmoid. Because of the non-linearity associated with this upper portion of the curve (see **Figure 6**), adult readers of greater perceptual ability produce attenuated effects of semantic priming for high frequency words, relative to the more additive effects observed among adults of low perceptual ability. By manipulating overall input strength associated with children and adults though the stimulus-onset asynchrony (SOA) of the prime, Plaut and Booth were able to induce interactive effects between semantic priming, word frequency, and perceptual ability in children by lengthening prime SOA, and more additive effects between these variables in adults by shortening SOA. Thus, Plaut and Booth's approach provides important theoretical insight into how a single mechanism (prime SOA) can yield a range of different behavioral outcomes for different individuals. However, without concrete specification of how the sigmoid maps onto the RT scale for the lexical decision task, connectionist models become unfalsifiable if the theory is able to simultaneously predict every form of

relationship between the factors, and the empirical data can be transformed by different parts of the sigmoidal function to produce any pattern of effect.

In general, we recommend against a "trial-and-error" approach to specification of the link function without firm theoretical guidance. However, such an approach might be considered if the statistical analysis has the truly exploratory goal of providing a description of how the dependent variable is affected by the predictors[5]. Critically, the focus of such exploratory analyses should not be on the statistical outcome of the fixed factors, because such tests assess how much the predictors affect the transformed metric rather than the dependent variable (Berry et al., 2010). Instead, the emphasis should center on how closely the description defined by the link function fits the observed data. Interestingly, the fit values determined by the AIC/BIC criteria favor the inverse link function over the identity link for all three experiments. Since we know of no current theory that explains why word frequency and stimulus quality are defined by an inverse relationship with RT, the fact that such a relationship is observed in the data remains of interest for future theoretical development.

Besides the mathematical form of the link function, we have also emphasized the importance of specifying an appropriate probability distribution for the dependent variable. Principally, this was achieved though theoretical consideration of the processes described by the probability distribution (e.g., RTs are more likely to reflect waiting time captured by a Gamma or Inverse Gaussian distribution than the number of times an event occurs in a Poisson distribution—even though the likelihood of observing extreme responses from both these processes are positively skewed). When multiple distributions provide equally plausible description of the processes underlying the dependent variable, as is the case with RT, the statistical analysis should be conducted using each of the distributions, with final selection based on the distribution that provides the closest fit to the observed data as determined by AIC/BIC fit statistics. Although the Inverse Gaussian distribution provided a superior fit for the experiments reported in Balota et al. (2013), the Gamma or other distributions not yet considered may provide a better match for other RT experiments.

Specifically, Rouder (2005) proposed that distributions for RT should also account for differences in minimum RT across experiments or individuals. Two-parameter distributions are ill-fitting because a third "shift" parameter is thought to be necessary in order to capture the fact that there is little or no mass below this minima in observed RTs. However, three-parameter Gamma or Inverse Gaussian distributions, which are similar to the shifted lognormal or shifted Weibull used by Ratcliff and Murdock (1976) and Rouder et al. (2008), are beyond the scope of GLMMs. This has led Rouder and colleagues to develop hierarchical models that use Bayesian statistics to make the necessary computations tractable (e.g., Rouder and Lu, 2005). Although such innovations will produce significant improvements over model fit as Bayesian techniques become better supported in popular statistical programs, the same careful consideration of the relationship between RT and the linear predictors (e.g., Rouder et al., 2008), and appreciation of models that capture rather than transform the attributes of RT are issues which remain pertinent for hierarchical Bayesian models.

While the results from the Balota et al. (2013) data suggest that better fitting distributions produce more precise standard errors and statistical greater power, the statistical outcomes from these datasets also seem to be relatively robust against moderate misspecification of the distribution in the GLMM framework. Given there is now evidence that experimental factors can produce isolated or even opposing effects on different parts of the RT distribution (e.g., Heathcote et al., 1991), GLMM analyses could be supplemented by consideration of how distributional shape is affected through variation in its parameters. An important step in this direction are the distributional analyses reported in Yap et al. (2009) that demonstrated differential effects of the experimental factors on the skewed tail of the RT distribution. By fitting ex-Gaussian distributions to the observed RTs, Yap et al. (2009) detected a significant four-way interaction between an individual's vocabulary ability, word frequency, non-word type and semantic priming on the τ parameter, reflecting stronger growth in the expression of semantic priming across the RT distribution for low compared to high frequency words particularly among those of lower vocabulary scores within a pseudo-homophone non-word environment. Importantly, transforming the data and analysing log or inverse RT would have obscured these findings of variation across individuals because the slowest condition - reflecting precisely those responses from low frequency words by those of poor vocabulary in a difficult pseudo-homophone non-word environment at the very tail of the distribution—would be more affected by the non-linear transformation than any of the other conditions (Balota et al., 2013). To extend these findings, future analyses could investigate these differences within the μ or λ parameters of the Inverse Gaussian distribution used in the present analyses, or to consider effects in three parameter distributions such as the ex-Gaussian or shifted Weibull (Rouder et al., 2008).

In summary, researchers are keenly aware of the potential biases associated with using skewed RT data for mean-based analyses. This has prompted recommendations to "transform away" these "erroneous…deviations from nature's ideals" (Speelman and McGann, 2013, p. 2), which exert even greater "undue influence" in skewed data than if responses had been normally distributed. By accommodating the shape of the skewed RT distribution, GLMMs remove the need to transform the dependent variable and allow the researcher to construct statistical models that answer their questions of interest, rather than being forced to change their question of interest to meet the constraints of the statistical model. Apart from alerting researchers to the problems associated with transforming their data and potentially obscuring systematic differences between individuals, the primary focus of this paper is to introduce an alternative solution and to describe the set of decisions required to correctly specify a GLMM. We have argued that the mental chronometry assumptions underlying much of the

[5]Other more appropriate methods, such as regression splines (Friedman and Roosen, 1995) and generalized additive models (Hastie and Tibshirani, 1990), are available if the goal is estimation of this relationship.

cognitive psychological research using RT data mean that the "correct metric" to analyse is often raw RT, but have illustrated scenarios for which transformed data might be more appropriate depending on the research question at hand. Should researchers have a clear theoretical basis for expecting a non-linear relationship between the predictors and the dependent variable, we have shown how specification of the form of the link function is able to achieve the same result in GLMMs without directly transforming the raw data. As the present analyses demonstrate, without such theoretical motivation, analyses based on non-linear transformations can lead researchers to spuriously conclude that an average effect is uniform across individuals or conditions (or vice versa) by altering the scale of the differences in an interaction to produce misleading or potentially contradictory results.

Acknowledgments

This research was supported by an Australian Postgraduate Award to SL and an Australian Research Council Discovery Project Grant DP120101491 to SA. The authors wish to thank David A. Balota and Melvin J. Yap for their generosity in supplying the datasets reported in Balota et al. (2013), and to R. Harald Baayen and Jeffery N. Rouder for constructive comments on an earlier version of this article.

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2015.01171

References

Andrews, S., and Heathcote, A. (2001). Distinguishing common and task-specific processes in word identification: a matter of some moment? *J. Exp. Psychol. Learn. Mem. Cogn.* 27, 514–544. doi: 10.1037/0278-7393.27.2.514

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics.* Cambridge: Cambridge University Press.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modelling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005

Balota, D. A., Aschenbrenner, A. J., and Yap, M. J. (2013). Additive effects of word frequency and stimulus quality: the influence of trial history and data transformations. *J. Exp. Psychol. Learn. Mem. Cogn.* 39, 1563–1571. doi: 10.1037/a0032186

Balota, D. A., and Yap, M. J. (2011). Moving beyond the mean in studies of mental chronometry the power of response time distributional analyses. *Curr. Dir. Psychol. Sci.* 20, 160–166. doi: 10.1177/0963721411408885

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2013). *lme4: Linear Mixed-effects Models using Eigen and S4. R Package Version 1.0-5.* Available online at: http://lme4.r-forge.r-project.org/

Berry, W. D., DeMeritt, J. H., and Esarey, J. (2010). Testing for interaction in binary logit and probit models: is a product term essential? *Am. J. Polit. Sci.* 54, 248–266. doi: 10.1111/j.1540-5907.2009.00429.x

Borowsky, R., and Besner, D. (1993). Visual word recognition: a multistage activation model. *J. Exp. Psychol. Learn. Mem. Cogn.* 19, 813–840. doi: 10.1037/0278-7393.19.4.813

Box, G. E. P., and Cox, D. R. (1964). An analysis of transformations. *J. R. Stat. Soc. Ser. B* 26, 211–252.

Brown, S., and Heathcote, A. J. (2008). The simplest complete model of choice reaction time: linear ballistic accumulation. *Cogn. Psychol.* 57, 153–178. doi: 10.1016/j.cogpsych.2007.12.002

Burnham, K. P., and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-theoretic Approach, 2nd Edn*. New York, NY: Springer-Verlag.

Carey, K. (2002). Hospital length of stay and cost: a multilevel modelling analysis. *Health Serv. Outcomes Res. Methodol.* 3, 41–56. doi: 10.1023/A:1021530924455

Carpenter, R. H. S., and Williams, M. L. L. (1995). Neural computation of log likelihood in control of saccadic eye movements. *Nature* 377, 59–62. doi: 10.1038/377059a0

Chapman, L. J., Chapman, J. P., Curran, T. E., and Miller, M. B. (1994). Do children and the elderly show heightened semantic priming? How to answer the question. *Dev. Rev.* 14, 159–185. doi: 10.1006/drev.1994.1007

Clark, H. H. (1973). The language-as-fixed-effect fallacy: a critique of language statistics in psychological research. *J. Verbal Learn. Verbal Behav.* 12, 335–359. doi: 10.1016/S0022-5371(73)80014-3

Cohen, J., Cohen, P., West, S. G., and Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioural Sciences, 3rd Edn*. Mahwah, NJ: Erlbaum.

Donders, F. (1868/1969). On the speed of mental processes. *Acta Psychol.* 30, 412–431. Transl. by W. G. Koster. doi: 10.1016/0001-6918(69)90065-1

Faust, M. E., Balota, D. A., Spieler, D. H., and Ferraro, F. R. (1999). Individual differences in information-processing rate and amount: implications for group differences in response latency. *Psychol. Bull.* 125, 777–799. doi: 10.1037/0033-2909.125.6.777

Forster, K. I., and Dickinson, R. G. (1976). More on the language-as-fixed-effect fallacy: Monte Carlo estimates of error rates for F 1, F 2, F', and min F'. *J. Verbal Learn. Verbal Behav.* 15, 135–142. doi: 10.1016/0022-5371(76)90014-1

Friedman, J. H., and Roosen, C. B. (1995). An introduction to multivariate adaptive regression splines. *Stat. Methods Med. Res.* 4, 197–217. doi: 10.1177/096228029500400303

Glass, G. V., Peckham, P. D., and Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying fixed effects analyses of variance and covariance. *Rev. Educ. Res.* 42, 237–288. doi: 10.3102/00346543042003237

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., and Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: the one- and two-factor fixed effects ANOVA cases. *J. Educ. Behav. Stat.* 17, 315–339.

Hasher, L., and Zacks, R. T. (1979). Automatic and effortful processes in memory. *J. Exp. Psychol. Gen.* 108, 356. doi: 10.1037/0096-3445.108.3.356

Hastie, T., and Tibshirani, R. (1990). *Generalized Additive Models.* London: Chapman and Hall.

Heathcote, A., Brown, S., and Cousineau, D. (2004). QMPE: estimating Lognormal, Wald, and Weibull RT distributions with a parameter-dependent lower bound. *Behav. Res. Methods Instrum. Comput.* 36, 277–290. doi: 10.3758/BF03195574

Heathcote, A., Popiel, S. J., and Mewhort, D. J. (1991). Analysis of response time distributions: an example using the Stroop task. *Psychol. Bull.* 109, 340. doi: 10.1037/0033-2909.109.2.340

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and their Applications*. New York, NY: Springer-Verlag.

Kinoshita, S., Mozer, M. C., and Forster, K. I. (2011). Dynamic adaptation to history of trial difficulty explains the effect of congruency proportion on masked priming. *J. Exp. Psychol. Gen.* 140, 622–636. doi: 10.1037/a0024230

Kliegl, R., Masson, M. E. J., and Richter, E. M. (2010). A linear mixed model analysis of masked repetition priming. *Vis. Cogn.* 18, 655–681. doi: 10.1080/13506280902986058

Lane, P. W. (2002). Generalized linear models in soil science. *Eur. J. Soil Sci.* 53, 241–251. doi: 10.1046/j.1365-2389.2002.00440.x

Laver, G. D., and Burke, D. M. (1993). Why do semantic priming effects increase in old age? A meta-analysis. *Psychol. Aging* 8, 34–43. doi: 10.1037/0882-7974.8.1.34

Lee, Y., Nelder, J. A., and Patiwan, Y. (2006). *Generalized Linear Models with Random Effects: Unified Analysis Via H-Likelihood*. Boca Raton, FL: Chapman and Hall.

Lix, L. M., Keselman, J. C., and Keselman, H. L. (1996). Consequences of assumption violations revisited: a quantitative review of alternatives to the one-way analysis of variance F test. *Rev. Educ. Res.* 66, 579–619.

Locker, L. Jr., Hoffman, L., and Bovaird, J. A. (2007). On the use of multilevel modeling as an alternative to items analysis in psycholinguistic research. *Behav. Res. Methods* 39, 723–730.

Loftus, G. R. (1978). On interpretation of interactions. *Mem. Cogn.* 6, 312–319. doi: 10.3758/BF03197461

Luce, R. D. (1986). *Response Times: Their Role in Inferring Elementary Mental Organization*. New York, NY: Oxford University Press.

Masson, M. E., and Kliegl, R. (2013). Modulation of additive and interactive effects in lexical decision by trial history. *J. Exp. Psychol. Learn. Mem. Cogn.* 39, 898–914. doi: 10.1037/a0029180

Matzke, D., and Wagenmakers, E. J. (2009). Psychological interpretation of the ex-Gaussian and shifted Wald parameters: a diffusion model analysis. *Psychon. Bull. Rev.* 16, 798–817. doi: 10.3758/PBR.16.5.798

McClelland, J. L., and Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: part 1. An account of basic findings. *Psychol. Rev.* 88, 375–407. doi: 10.1037/0033-295X.88.5.375

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models, 2nd Edn*. New York, NY: Chapman and Hall.

Morton, J. (1969). Interaction of information in word recognition. *Psychol. Rev.* 76, 165. doi: 10.1037/h0027366

Myerson, J., Ferraro, F. R., Hale, S., and Lima, S. D. (1992). General slowing in semantic priming and word recognition. *Psychol. Aging* 7, 257–270. doi: 10.1037/0882-7974.7.2.257

Pinheiro, J. C., and Bates, D. M. (2000). *Mixed-effects Models in S and S-PLUS.* New York, NY: Springer.

Plaut, D. C., and Booth, J. R. (2000). Individual and developmental differences in semantic priming: empirical and computational support for a single-mechanism account of lexical processing. *Psychol. Rev.* 107, 786–823. doi: 10.1037/0033-295X.107.4.786

Posner, M. I. (1978). *Chronometric Explorations of Mind*. Oxford: Lawrence Erlbaum.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychol. Rev.* 85, 59–108. doi: 10.1037/0033-295X.85.2.59

Ratcliff, R., and Murdock, B. B. (1976). Retrieval processes in recognition memory. *Psychol. Rev.* 83, 190–214. doi: 10.1037/0033-295X.83.3.190

Ratcliff, R., and Rouder, J. N. (2000). A diffusion model account of masking in two-choice letter identification. *J. Exp. Psychol. Hum. Percept. Perform.* 26, 127–140. doi: 10.1037/0096-1523.26.1.127

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods, 2nd Edn.* Thousand Oaks, CA: Sage Press.

R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: http://www.R-project.org

Rouder, J. N. (2005). Are unshifted distributional models appropriate for response time? *Psychometrika* 70, 377–381. doi: 10.1007/s11336-005-1297-7

Rouder, J. N., and Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychon. Bull. Rev.* 12, 573–604. doi: 10.3758/BF03196750

Rouder, J. N., Tuerlinckx, F., Speckman, P., Lu, J., and Gomez, P. (2008). A hierarchical approach for fitting curves to response time measurements. *Psychon. Bull. Rev.* 15, 1201–1208. doi: 10.3758/PBR.15.6.1201

Salthouse, T. A. (1985). "Speed of behavior and its implications for cognition," in *Handbook of the Psychology of Aging, 2nd Edn,* eds J. E. Birren and K. W. Schaie (New York, NY: Van Nostrand Reinhold), 400–426.

Schwarz, W. (2001). The ex-Wald distribution as a descriptive model of response times. *Behav. Res. Methods Instrum. Comput.* 33, 457–469. doi: 10.3758/BF03195403

Speelman, C. P., and McGann, M. (2013). How mean is the mean? *Front. Psychol.* 4:451. doi: 10.3389/fpsyg.2013.00451

Stanners, R. F., Jastrzembski, J. E., and Westbrook, A. (1975). Frequency and visual quality in a word-nonword classification task. *J. Verbal Learn. Verbal Behav.* 14, 259–264. doi: 10.1016/S0022-5371(75)80069-7

Sternberg, S. (1969). Memory-scanning: mental processes revealed by reaction-time experiments. *Am. Sci.* 57, 421–457.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science* 103, 677–680. doi: 10.1126/science.103.2684.677

Stroup, W. W. (2013). *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. Boca Raton, FL: CRC Press.

Townsend, J. T. (1992). "On the proper scales of reaction time," in *Cognition, Information Processing, and Psychophysics: Basic Issues,* eds H. G. Geissler, S. W. Link, and J. T. Townsend (Hillsdale, NJ: Erlbaum), 105–120.

Trafimow, D. (2014). The mean as a multilevel issue. *Front. Psychol.* 5:180. doi: 10.3389/fpsyg.2014.00180

Van Zandt, T., and Ratcliff, R. (1995). Statistical mimicking of reaction time data: single-process models, parameter variability, and mixtures. *Psychon. Bull. Rev.* 2, 20–54. doi: 10.3758/BF03214411

Wagenmakers, E. J., and Brown, S. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychol. Rev.* 114, 830–841. doi: 10.1037/0033-295X.114.3.830

Yap, M. J., and Balota, D. A. (2007). Additive and interactive effects on response time distributions in visual word recognition. *J. Exp. Psychol. Learn. Mem. Cogn.* 33, 274–296. doi: 10.1037/0278-7393.33.2.274

Yap, M. J., Balota, D. A., Tse, C. S., and Besner, D. (2008). On the additive effects of stimulus quality and word frequency in lexical decision: evidence of opposing interactive influences revealed by RT distributional analyses. *J. Exp. Psychol. Learn. Mem. Cogn.* 34, 495–513. doi: 10.1037/0278-7393.34.3.495

Yap, M. J., Tse, C. S., and Balota, D. A. (2009). Individual differences in the joint effects of semantic priming and word frequency revealed by RT distributional analyses: the role of lexical integrity. *J. Mem. Lang.* 61, 303–325. doi: 10.1016/j.jml.2009.07.001

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Lo and Andrews. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 06 January 2015
doi: 10.3389/fpsyg.2014.01492

To center or not to center? Investigating inertia with a multilevel autoregressive model

Ellen L. Hamaker[1]* and Raoul P. P. P. Grasman[2]

[1] Methodology and Statistics, Faculty of Social and Behavioural Sciences, Utrecht University, Utrecht, Netherlands
[2] Psychological Methods, University of Amsterdam, Amsterdam, Netherlands

Edited by:
Marek McGann, Mary Immaculate College, Ireland

Reviewed by:
Pietro Cipresso, IRCCS Istituto Auxologico Italiano, Italy
Johnny Zhang, University of Notre Dame, USA

***Correspondence:**
Ellen L. Hamaker, Methodology and Statistics, Faculty of Social and Behavioural Sciences, Utrecht University, PO Box 80140, 3508 TC, Utrecht, Netherlands
e-mail: e.l.hamaker@uu.nl

Whether level 1 predictors should be centered per cluster has received considerable attention in the multilevel literature. While most agree that there is no one preferred approach, it has also been argued that cluster mean centering is desirable when the within-cluster slope and the between-cluster slope are expected to deviate, and the main interest is in the within-cluster slope. However, we show in a series of simulations that if one has a multilevel autoregressive model in which the level 1 predictor is the lagged outcome variable (i.e., the outcome variable at the previous occasion), cluster mean centering will in general lead to a downward bias in the parameter estimate of the within-cluster slope (i.e., the autoregressive relationship). This is particularly relevant if the main question is whether there is on average an autoregressive effect. Nonetheless, we show that if the main interest is in estimating the effect of a level 2 predictor on the autoregressive parameter (i.e., a cross-level interaction), cluster mean centering should be preferred over other forms of centering. Hence, researchers should be clear on what is considered the main goal of their study, and base their choice of centering method on this when using a multilevel autoregressive model.

Keywords: centering, autoregressive models, multilevel models, dynamics, inertia

Longitudinal data are characterized by a nested structure, in which occasions are clustered within individuals. While such data are traditionally analyzed using repeated measures ANOVA, this approach is restrictive in that it requires an equal number of observations for each participant. A further limitation associated with repeated measures ANOVA is that the results pertain to the aggregate and may not be meaningful for any particular individual. A more sophisticated approach—which overcomes these limitations—is *multilevel modeling* (Singer and Willett, 2003; Hox, 2010; Snijders and Bosker, 2012; also known as mixed modeling, see Verbeke and Molenberghs, 2000; hierarchical modeling, see Raudenbush and Bryk, 2002; or random-effects modeling, see Laird and Ware, 1982): This approach can be used for (highly) unbalanced longitudinal data, and it allows for individual trajectories over time. The latter implies we can study between-person (or interindividual) differences in within-person (or intraindividual) patterns of change.

It is not uncommon for the residuals of a longitudinal multilevel model to be autocorrelated, meaning that residuals are related to each other over time. Failing to account for this may bias the estimates of the standard errors, and as a result affect the inferences based on them. Therefore, multilevel software packages include the option to control for autocorrelation through specifying diverse structures for the errors, such as a Toeplitz matrix, or a first order autoregressive process. Alternatively, autocorrelation can be modeled explicitly through the inclusion of the lagged outcome variable (that is, the outcome variable at the previous occasion) as a covariate. Such models have been referred to as (prospective) change models (e.g., Larson and Almeida, 1999), and are used to investigate the—potentially causal—effect of a (lagged) predictor on the outcome variable, while "controlling" or "adjusting" for the previous level of the outcome variable (e.g., Bolger and Zuckerman, 1995; Gunthert et al., 2007; Moberly and Watkins, 2008; Henquet et al., 2010).

While autocorrelation is typically considered a nuisance in longitudinal multilevel modeling, there are a few multilevel studies that focus specifically on the autoregressive relationship between consecutive observations, and on individual differences therein (cf., Suls et al., 1998; Rovine and Walls, 2006; Kuppens et al., 2010; Koval and Kuppens, 2012; Wang et al., 2012; Brose et al., 2014). The interest in an individual's autoregressive parameter comes from the fact that this parameter is related to the time it takes the individual to recover from a perturbation and restore equilibrium: While an autoregressive parameter close to zero implies that there is little carryover from one measurement occasion to the next and recovery is thus instant, an autoregressive parameter close to one implies that there is considerable carryover between consecutive measurement occasions, such that perturbations continue to have an effect on subsequent occasions. For this reason, the autoregressive parameter can also be considered as a measure of *inertia* or *regulatory weakness*.

Empirical studies have shown that individual differences in inertia in emotions and affect are positively related to neuroticism and depression, in that people higher on neuroticism or depression take longer to restore equilibrium than others (Suls et al., 1998; Kuppens et al., 2010; Wang et al., 2012). In addition, women tend to have higher inertia than men in both their daily affect (Wang et al., 2012), and their daily drinking behavior (Rovine and

Walls, 2006). In a prospective study, Kuppens et al. (2012) showed that affective inertia at age 9–12 was predictive of the onset of depression two and a half years latter, corresponding to the idea that high inertia is reflective of a maladaptive regulation mechanism. Similarly, Wang et al. (2012) showed that inertia is positively related to later detrimental health outcomes. Furthermore, inertia has been shown to be related—but not identical—to rumination (Koval et al., 2012) and perseverative thoughts (Brose et al., 2014), and is positively related to depression even after these related characteristics are taken into account. Taken together, these studies show that inertia is a meaningful individual characteristic that is reflective of a maladaptive regulatory mechanism that is associated with both current *and* future well-being.

To model individual differences in inertia, the above studies all relied on multilevel modeling based on a first-order autoregressive process: In this model, the level 1 predictor is formed by the lagged outcome variable, and its random slope thus represents individual differences in inertia. A pressing question in this context is whether the autoregressive predictor should be centered per person or not. This is a rather fundamental issue, as it is well-known from the multilevel literature that the centering method used for a level 1 predictor (i.e., no centering, centering with the grand mean, or centering per cluster), affects the results (cf. Kreft et al., 1995; Raudenbush and Bryk, 2002; Hox, 2010; Snijders and Bosker, 2012). The consensus seems to be that there is no one preferred method and that the choice should depend on the specific situation and the research question (cf. Kreft et al., 1995; Nezlek, 2001; Snijders and Bosker, 2012). One such specific situation is described by Raudenbush and Bryk (2002), who indicate that if the within-cluster and between-cluster slopes differ, centering per cluster should be preferred, because failing to do so will lead to results that are "uninterpretable" (p. 135). Furthermore, Enders and Tofighi (2007) argue that if there is a clear interest in the within-cluster slope, centering per cluster is recommendable.

With this latter advice in mind, centering the lagged autoregressive predictor per person seems the right approach, because: (a) we are interested in the within-person slope; and (b) we expect the within-person slope to differ from the between-person slope.[1] The aim of the current paper is therefore to investigate whether the advice formulated by Raudenbush and Bryk (2002) and Enders and Tofighi (2007) also applies to the multilevel autoregressive model with a random slope that represents individual differences in inertia. To this end, we begin by presenting the multilevel autoregressive model and discuss its interpretation. To make the model compatible with standard multilevel software, we discuss two parameterizations—based on different centering methods—and we show through an empirical application that these lead to different results for the inertia parameter. In the second section we draw from several key publications in the multilevel literature and discuss the effects of centering a level 1 predictor. The third section contains simulations based on the standard multilevel model to verify some of the claims made in the literature. Additionally, we simulate the multilevel autoregressive model to investigate how centering affects the estimation of inertia. In the fourth section we apply the insights obtained from the simulation study to the empirical data set. We end by presenting recommendations to the researcher interested in studying inertia using the multilevel autoregressive model, either with or without level 2 predictors.

1. MULTILEVEL AUTOREGRESSIVE MODEL

Many applications of longitudinal multilevel modeling consist of modeling deterministic trajectories over time, for instance a linear or quadratic trend. While such models are extremely useful for studying developmental processes (cf., Curran and Bauer, 2011), they may be less useful when the longitudinal data comprise daily affective or symptom measurements, or affective ratings in an observation study: Then the interest may be not so much in overall trends (as they are likely to be absent from the data), but rather in the dynamics of a stationary process, that is, a process that is characterized by changes *over time*, while these changes are not directly a *function of time*. A promising model for this purpose is the multilevel autoregressive model, which has been successfully applied in an increasing number of studies (e.g., Suls et al., 1998; Rovine and Walls, 2006; Kuppens et al., 2010; Koval and Kuppens, 2012; Wang et al., 2012; Brose et al., 2014).

We begin this section by presenting the multilevel autoregressive model using a parametrization that we consider to be most useful from a substantive viewpoint. However, since this parametrization is not compatible with standard multilevel software, we also present two alternative parametrization of this model, and discuss their advantages and disadvantages. We apply both parameterizations to an empirical data set consisting of daily measurement of positive and negative affect.

1.1. A MODEL TO STUDY INDIVIDUAL DIFFERENCES IN MEAN AND INERTIA

Let y_{ti} be the observation for individual i at occasion t, for instance the person's negative affect or self-esteem measured at a daily basis, with $i = 1, \ldots, N$ and $t = 1, \ldots, T_i$. The most basic model for such nested data would be a model which allows for individual differences in means. At level 1 the observations are then modeled as

$$y_{ti} = \mu_i + a_{ti} \tag{1}$$

where μ_i represents the individual's mean score, which can be interpreted as his/her trait score or equilibrium, while a_{ti} is the individual's temporal deviation from this equilibrium; and at level 2 the individual means are then modeled as

$$\mu_i = \mu + u_{0i} \tag{2}$$

where μ is the grand mean, and u_{0i} is the individual's deviation from the grand mean. These deviations are assumed to be normally distributed, that is, $u_{0i} \sim N(0, \sigma^2_{u0})$.[2]

[1] As we will show later on, the between-person slope will always be (essentially) 1 in this model, while the within-person slope is expected to lie between −1 and 1.

[2] Note that the model presented in Equations 1 and 2 corresponds to what is known as a random intercept model or empty model, and is typically considered as one of the options in longitudinal multilevel modeling.

If repeated measures are taken (relatively) close in time, the current measurement is likely to be predictable from the preceding measurement. That is, the individual's deviation from his/her equilibrium at a particular occasion is likely to affect the deviation at the next occasion, which can be expressed as a first order autoregressive model, that is,

$$a_{ti} = \phi_i a_{t-1,i} + e_{ti} \tag{3}$$

where the residuals e_{ti} are independently and identically distributed, with $e_{ti} \sim N(0, \sigma_e^2)$. This residual e_{ti} can be thought of as representing everything that influences the process under investigation: For instance, if we are measuring negative affect, factors that are likely to influence this process include the occurrence of negative or stressful events, the appraisal of these events and the associations and memories that they trigger, but also psychophysiological factors like caffeine or alcohol consumption, et cetera.

The autoregressive parameter ϕ_i relates the outcome variable to itself at the preceding occasion, and thus represents the *inertia* of the person. For an autoregressive process to be stationary, the autoregressive parameter has to lie between -1 and 1 (e.g., Hamilton, 1994). Note however that this does not imply that the autoregressive parameter is truly restricted to this range: Values larger than 1 (or smaller than -1) are possible, but the resulting process would no longer be a stationary process. In psychological research, this parameter typically lies somewhere between 0 and 0.6 (e.g., Rovine and Walls, 2006; Wang et al., 2012), and we are therefore not concerned about boundary constraints when estimating this model.

The individual differences in the autoregressive parameter can be modeled at level 2 as

$$\phi_i = \phi + u_{1i} \tag{4}$$

where ϕ denotes the average autoregressive parameter across people, and u_{1i} denotes the individual's deviation from this average, with $u_{1i} \sim N(0, \sigma_{u1}^2)$. Furthermore, the individuals' means and their autoregressive parameters may be correlated, as represented by the covariance between u_{0i} and u_{1i}, which is denoted as $\sigma_{u0,u1}$. Wang et al. (2012) for instance found a significant positive correlation of 0.40 between the individuals' means μ_i and their autoregressive parameters ϕ_i based on daily measurements of negative affect.

1.2. MAKING THE MODEL COMPATIBLE WITH STANDARD MULTILEVEL SOFTWARE

The model in Equations 1–4 represents the multilevel autoregressive model, where Equations 1 and 3 form level 1, while Equations 2 and 4 form level 2. However, most multilevel software packages do not allow for formulating a level 1 model using more than one equation. We consider two solutions for this.

The first solution consists of specifying the model at level 1 as

$$y_{ti} = c_i + \phi_i^n y_{t-1,i} + e_{ti}, \tag{5}$$

and at level 2 as

$$c_i = \gamma_{00}^n + u_{0i}^n \tag{6}$$

$$\phi_i^n = \gamma_{10}^n + u_{1i}^n, \tag{7}$$

where the superscript n indicates that in this approach *no centering* (NC) was used (i.e., the raw data were used). The relationship between the model specified in Equations 1–4 and the model specified in Equations 5–7 is shown in Appendix 1; however, while the multilevel models presented here are *structurally* the same, the current formulation is based on the assumption that c_i is normally distributed, which necessarily implies that μ_i will not have a normal distribution (as it is a function of c_i and ϕ_i, see Appendix 1). This is detrimental, as we are typically interested in μ_i as representing an individual's average or trait score, and assume these trait scores to be normally distributed in the population. In contrast, c_i is a rather arbitrary score (i.e., the expected score when the individual scored zero on the preceding occasion), that is of limited (or no) substantive interest, and for which we do not have a particular distributional expectation. Also, if we are interested in including predictors at level 2, we would prefer to use these as predictors of μ_i, rather than of c_i.

Therefore, we consider a second solution, which is based on using the individually centered lagged autoregressive predictor $(y_{t-1,i} - \mu_i)$, such that the model at level 1 is

$$y_{ti} = \mu_i + \phi_i^c (y_{t-1,i} - \mu_i) + e_{ti} \tag{8}$$

and at level 2 it is

$$\mu_i = \gamma_{00}^c + u_{0i}^c \tag{9}$$

$$\phi_i^c = \gamma_{10}^c + u_{1i}^c \tag{10}$$

where the superscript c implies that the level 1 predictor was subjected to *cluster mean centering* (CMC; also referred to as within-group or within-person centering). The advantage of the current approach over the previous one is that it results in μ_i and ϕ_i being the random coefficients that are subsequently modeled at level 2. However, it also presents us with a catch-22: To center the lagged predictor, we need an estimate of μ_i, which we actually need to estimate using this model. We will consider several solutions to this problem in our simulations, including the use of the sample mean per person.

1.3. APPLICATION: PART 1

To investigate whether the two approaches proposed above lead to the same or different results for the inertia parameter, we apply the two parameterizations of the multilevel autoregressive model to an empirical data set that was obtained as part of the Dynamics of Dyadic Interactions Project at the University of California, Davis (Ferrer and Widaman, 2008; Ferrer et al., 2012). The data used here consist of daily measurements of relationship specific positive and negative affect. We analyzed these data for men and women separately (sample sizes 193 and 192, respectively), using multilevel autoregressive models with random intercepts (i.e., c_i, based on NC) or means (i.e., μ_i, based on CMC), and random autoregressive parameters (ϕ_i). The estimates for the fixed effects parameters γ_{00} and γ_{10} are presented in **Table 1**.

Table 1 | Results for multilevel autoregressive model (with random effects).

			NC		CMC	
Males	PA	γ_{00}	2.167	[2.044, 2.290]	3.518	[3.425, 3.611]
		γ_{10}	0.387	[0.357, 0.417]	0.353	[0.322, 0.384]
	NA	γ_{00}	0.971	[0.923, 1.020]	1.344	[1.300, 1.389]
		γ_{10}	0.268	[0.235, 0.301]	0.242	[0.208, 0.275]
Females	PA	γ_{00}	2.220	[2.095, 2.346]	3.491	[3.392, 3.590]
		γ_{10}	0.370	[0.340, 0.399]	0.341	[0.311, 0.370]
	NA	γ_{00}	0.978	[0.935, 1.021]	1.348	[1.304, 1.392]
		γ_{10}	0.255	[0.222, 0.288]	0.225	[0.192, 0.258]

Estimates for the fixed effects parameters in a multilevel autoregressive model (with random intercept and slope). The 95% confidence intervals are given between brackets. Estimation was based on using NC or CMC (with the sample means) for the lagged autoregressive predictor. Fixed effects are: (a) γ_{00}, which represents the averaged intercept when using NC, or the grand mean when using CMC; and (b) γ_{10}, which represents the averaged (i.e., fixed effects) autoregressive parameter.

It shows that the parameter estimates obtained with the two models are not identical. This is not surprising as we are already aware that the two parameterizations differ with respect to the meaning of γ_{00}. However, it also shows that the parameter estimates for γ_{10}—which represents the average inertia in both parameterizations—differ from each other. Especially when considering relationship specific PA in males, it can be seen that CMC and NC lead to estimates of the inertia that are not covered by the 95% confidence interval of the alternative parametrization (implying these estimates are relatively different).

The question thus arises, which approach should be preferred—NC or CMC—when the interest is in obtaining an appropriate estimate of the average autoregressive parameter. As this touches upon the more general topic of whether level 1 predictors should be centered or not in multilevel models, we first consult the multilevel literature with respect to centering level 1 predictors.

2. TO CENTER OR NOT TO CENTER: A PERSISTING QUESTION IN MULTILEVEL MODELING

Centering a level 1 predictor in multilevel modeling is a complicated affaire. While there are several sources that provide excellent coverage of this topic (e.g., Kreft et al., 1995; Snijders and Bosker, 2012), it still seems to create much confusion, especially amongst the more novice users. A fundamental issue when dealing with a level 1 predictor is the fact that the relationship between a predictor and an outcome variable may differ across levels. For instance, consider the hypothetical example in the left panel of **Figure 1**, representing the relationship between typing speed and number of typos. This relationship is likely to be positive within individuals (i.e., at level 1), in that a person tends to make more mistakes if he/she types faster. However, the relationship across individuals (i.e., at level 2) is likely to be negative, because individuals who tend to type fast *on average*, also tend to be more experienced and therefore make fewer mistakes *on average* (cf. Hamaker, 2012; see also Nezlek, 2001; Enders and Tofighi, 2007; Kievit et al., 2013).

In this section we discuss the effects of different centering methods, when there are different slopes at the two levels. Our main interest is in obtaining an appropriate estimate for the within-cluster slope, as this is most informative with respect to the within-person process. To facilitate the transition to the multilevel autoregressive model, we will present the issue based on repeated measures within individuals (rather than individuals organized in groups). In following Raudenbush and Bryk (2002) and Enders and Tofighi (2007), we begin by considering models with a fixed slope only. Subsequently, we discuss contextual models, in which the cluster means are included as a predictor at level 2. Then we discuss extensions that allow for random slopes. We end this section by speculating on the effects of centering in the context of the multilevel autoregressive model.

2.1. THE WITHIN-CLUSTER AND BETWEEN-CLUSTER SLOPES IN MULTILEVEL DATA WITH A FIXED SLOPE

Suppose that x_{ti} is the predictor, such as typing speed or the occurrence of a negative event, and that y_{ti} is the outcome variable, such as number of typos or negative affect. Let $i = 1, \ldots, N$ denote the individual, and $t = 1, \ldots, T_i$ denote the measurement occasion within individual i. Raudenbush and Bryk (2002) discuss how to obtain estimates of the between-person slope, relating the trait scores on the outcome variable to the trait scores on the predictor, and of the averaged or pooled within-person slope, describing the process that operates within individuals, using ordinary least squares (OLS). To this end, we first need the individual means on the predictor and the outcome variable, that is,

$$\bar{x}_{\cdot i} = \frac{1}{T_i}\sum_{i=1}^{T_i} x_{ti} \quad \text{and} \quad \bar{y}_{\cdot i} = \frac{1}{T_i}\sum_{i=1}^{T_i} y_{ti}. \tag{11}$$

Then the between-person or *between-cluster* slope β_B can be obtained by analyzing these individual means using the regression equation

$$\bar{y}_{\cdot i} = \beta_0 + \beta_B \bar{x}_{\cdot i} + e_i. \tag{12}$$

Additionally, the averaged within-person or *within-cluster* slope β_W can be obtained through applying CMC to both the predictor and the outcome variable, and analyze these data for individuals simultaneously, that is

$$y_{ti} - \bar{y}_{\cdot i} = \beta_W\left(x_{ti} - \bar{x}_{\cdot i}\right) + e_{ti}. \tag{13}$$

Clearly, β_W and β_B need not be the same.

Raudenbush and Bryk (2002) discuss how the slopes from diverse multilevel approaches are related to these two basic slopes. There are three approaches that can be used, that is: no centering (NC), grand-mean centering (GMC), and cluster-mean centering. GMC is simply a linear transformation of the data, and leads to a model that is statistically equivalent to NC (cf. Kreft et al., 1995; Raudenbush and Bryk, 2002; Snijders and Bosker, 2012). Therefore, we do not discuss this approach separately, and only focus on the comparison between NC and CMC below.

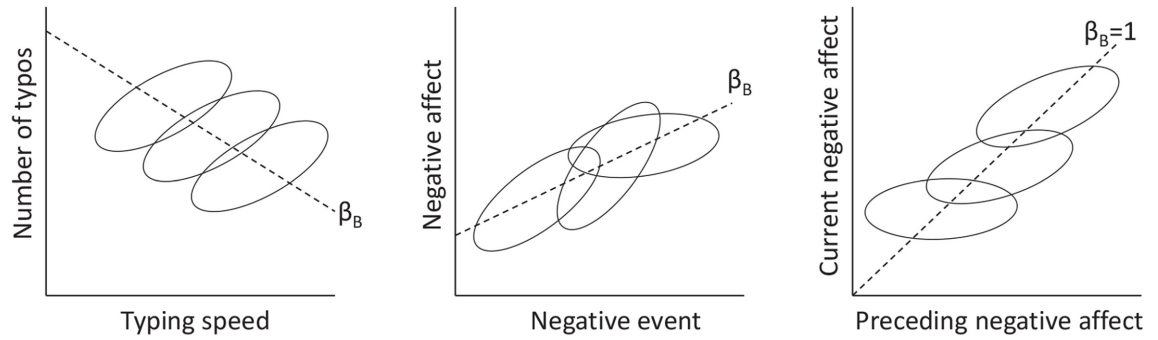


FIGURE 1 | Illustration of within-person and between-person relationships between two variables. Each ellipse represents the data from a single person. Dashed lines represent the between-person slope (i.e., β_B), which may have a different sign as the within-person slope (**Left panel**), may be similar to the (average or fixed) within-person slope (**Middle panel**), or may be larger than the (average or fixed) within-person slope (**Right panel**).

The model based on NC—with a random intercept and a fixed slope—can be expressed as

$$\begin{aligned} y_{ti} &= \alpha_i^n + \beta_i^n x_{ti} + e_{ti} \\ \alpha_i^n &= \gamma_{00}^n + u_{0i}^n \\ \beta_i^n &= \gamma_{10}^n. \end{aligned} \tag{14}$$

whereas the corresponding model based on CMC of the predictor results in

$$\begin{aligned} y_{ti} &= \alpha_i^c + \beta_i^c\left(x_{ti} - \bar{x}_{\cdot i}\right) + e_{ti} \\ \alpha_i^c &= \gamma_{00}^c + u_{0i}^c \\ \beta_i^c &= \gamma_{10}^c. \end{aligned} \tag{15}$$

Note that the fixed slope γ_{10}^c from the CMC model is analogous to β_W in Equation 13, with $y_{ti} - \bar{y}_{\cdot i}$ being replaced by $y_{ti} - \alpha_i^c$. Hence, CMC leads to an estimate of the within-cluster slope. The question is whether the within-cluster slope can also be obtained from the model in Equation 14.

To this end, we enter the level 2 expressions into the level 1 expressions, such that the model based on NC can be expressed as

$$y_{ti} = \gamma_{00}^n + \gamma_{10}^n x_{ti} + u_{0i}^n + e_{ti}, \tag{16}$$

and the model based on CMC can be expressed as

$$y_{ti} = \gamma_{00}^c + \gamma_{10}^c x_{ti} - \gamma_{10}^c \bar{x}_{\cdot i} + u_{0i}^c + e_{ti}. \tag{17}$$

From these expressions it becomes clear that these models are not equivalent, as one cannot be considered an alternative parametrization of the other (cf. Kreft et al., 1995). This also implies that the within-cluster slope cannot be derived based on the results obtained from NC. Raudenbush and Bryk (2002) indicate that the slope of the level 1 predictor obtained with NC (γ_{10}^n) is "an uninterpretable blend" (p. 139) of the averaged within-cluster slope β_W and the between-cluster slope β_B. This has led them to formulate the advice to use CMC whenever the interest is in obtaining an unbiased estimate of the within-cluster relationship, and the within-cluster and between-cluster relationships are expected to differ from each other (see also Enders and Tofighi, 2007).[3]

One could argue that the models above are not correct, because the between-cluster relationship is not explicitly modeled. Raudenbush and Bryk (2002) discuss the option of obtaining estimates of both β_W and β_B in a single multilevel model, through including the cluster means on the predictor as a level 2 predictor for the intercept. In case of NC, this results in

$$\begin{aligned} y_{ti} &= \alpha_i^n + \beta_i^n x_{ti} + e_{ti} \\ \alpha_i^n &= \gamma_{00}^n + \gamma_{01}^n \bar{x}_{\cdot i} + u_{0i}^n, \\ \beta_i^n &= \gamma_{10}^n. \end{aligned} \tag{18}$$

while in case of CMC this gives

$$\begin{aligned} y_{ti} &= \alpha_i^c + \beta_i^c\left(x_{ti} - \bar{x}_{\cdot i}\right) + e_{ti} \\ \alpha_i^c &= \gamma_{00}^c + \gamma_{01}^c \bar{x}_{\cdot i} + u_{0i}^c, \\ \beta_i^c &= \gamma_{10}^c. \end{aligned} \tag{19}$$

In the latter approach, the within-cluster slope is again represented by γ_{10}^c, and now the between-cluster slope is represented by γ_{01}^c.

To see whether NC and CMC lead to equivalent models in this case, we substitute the level 2 expressions in the level 1 expression. For NC this results in

$$y_{ti} = \gamma_{00}^n + \gamma_{01}^n \bar{x}_{\cdot i} + \gamma_{10}^n x_{ti} + u_{0i}^n + e_{ti}, \tag{20}$$

[3]Note that there may be situations in which the within-person slope and the between-person slope do not differ that much, such that failing to separate them does not affect the results very much: For instance, in the middle panel of **Figure 1** the relationship between negative events and negative affect is represented, which shows that within individuals, there is (on average) a positive relationship (i.e., people tend to experience more negative affect on days that more negative events occur), and the between-person relationship is very similar (i.e., people who tend to experience more negative events on average, also tend to have higher levels of negative affect on average). In this case, using CMC or NC/GMC will not change the estimate of the fixed slope very much. However, the point remains that to obtain an adequate estimate of the averaged within-cluster slope, CMC should be preferred.

and for CMC it results in

$$y_{ti} = \gamma^c_{00} + \left(\gamma^c_{01} - \gamma^c_{10}\right)\bar{x}_{\cdot i} + \gamma^c_{10}x_{ti} + u^c_{0i} + e_{ti}, \qquad (21)$$

showing the models are equivalent. Furthermore, it becomes clear that actually both models provide an estimate of the within-cluster slope, that is, $\gamma^n_{10} = \gamma^c_{10} = \beta_W$. Additionally, we have $\gamma^n_{01} = \gamma^c_{01} - \gamma^c_{10} = \beta_B - \beta_W$, that is, γ^n_{01} represents the *difference* in the between-cluster and the within-cluster slopes. This is also referred to as the *contextual* or *compositional effect* (cf. Raudenbush and Bryk, 2002 p. 141).[4], and these models are referred to as contextual models.

In sum, when there is a fixed slope, NC, and CMC lead to equivalent models if one includes the cluster means for the level 1 predictor as a level 2 predictor for the intercept (Kreft et al., 1995). However, if the cluster means are not included, these two models are not equivalent.

2.2. EFFECTS OF NC AND CMC WHEN THERE IS A RANDOM SLOPE

While the model equivalence above is interesting, it is of limited value in practice, as we are often interested in models with random slopes. For instance, consider the middle panel of **Figure 1**, representing the hypothetical relationship between the number of negative events and negative affect in daily measurements: It shows that the strength of the within-person relationship differs across individuals.

Snijders and Bosker (2012) show that if there is a random slope, the model equivalence presented above no longer holds. Allowing for a random slope in the NC model in Equation 18, implies we have $\beta^n_i = \gamma^n_{10} + u^n_{1i}$, and we can thus write

$$y_{ti} = \gamma^n_{00} + \gamma^n_{01}\bar{x}_{\cdot i} + \gamma^n_{10}x_{ti} + u^n_{1i}x_{ti} + u^n_{0i} + e_{ti}. \qquad (22)$$

For the CMC model in Equation 19, a random slope implies we have $\beta^c_i = \gamma^c_{10} + u^c_{1i}$, such that the model can be expressed as

$$\begin{aligned} y_{ti} = {} & \gamma^c_{00} + \left(\gamma^c_{01} - \gamma^c_{10}\right)\bar{x}_{\cdot i} + \gamma^c_{10}x_{ti} + u^c_{1i}x_{ti} \\ & - u^c_{1i}\bar{x}_{\cdot i} + u^c_{0i} + e_{ti}. \end{aligned} \qquad (23)$$

This shows that—once there is a random slope—these models are no longer statistically equivalent, as they differ with respect to the term $\left(-u^c_{1i}\bar{x}_{\cdot i}\right)$. However, Kreft et al. (1995) pointed out that the fixed effect within-cluster slope is still the same across these two models: That is, $\gamma^n_{10} = \gamma^c_{10} = \beta_W$ (see Kreft et al., 1995, p. 13). Hence, when the goal is to obtain an estimate of the within-cluster slope, and the within-cluster and between-cluster slope are expected to differ, it seems that one can chose either use CMC, or the contextual versions of CMC or NC/GMC: Although the contextual models are not equivalent when a random slope is included, they will result in the same within-cluster slope estimate.

For the sake of completeness, we also provide the expression for the models that include a random slope but without the cluster means as a level 2 predictor—as these are more common than the contextual models and the fixed slope models discussed above. In that case, NC leads to

$$y_{ti} = \gamma^n_{00} + \gamma^n_{10}x_{ti} + u^n_{1i}x_{ti} + u^n_{0i} + e_{ti}. \qquad (24)$$

and CMC leads to

$$y_{ti} = \gamma^c_{00} + \gamma^c_{10}x_{ti} - \gamma^c_{10}\bar{x}_{\cdot i} + u^c_{1i}x_{ti} - u^c_{1i}\bar{x}_{\cdot i} + u^c_{0i} + e_{ti}. \qquad (25)$$

As expected based on what was discussed above, both the fixed and the random parts of these models differ, and only CMC leads to an estimate of the average within-cluster slope, while NC leads to a slope that represents some mix of the within-cluster and between-cluster slopes.

2.3. PRELIMINARY THOUGHTS ON CENTERING IN THE MULTILEVEL AUTOREGRESSIVE MODEL

Before turning to our simulation study, we speculate briefly on the effects of NC and CMC in case of the multilevel autoregressive model. The contextual model would imply that we include the within-person means as a predictor for the intercept at level 2, that is

$$\begin{aligned} y_{ti} &= \mu_i + \phi_i\left(y_{t-1,i} - \mu_i\right) + e_{ti} \\ \mu_i &= \gamma_{00} + \gamma_{01}\mu_i + u_{0i} \\ \phi_i &= \phi + u_{1i}, \end{aligned} \qquad (26)$$

where $\phi = \beta_W$ is the average within-cluster relationship, and $\gamma_{01} = \beta_B$ is the between-cluster relationship. Note however that now μ_i appears on both sides of the equality sign, and it follows that $\gamma_{01} = 1$, $\gamma_{00} = 0$ and $u_{0i} = 0$ (and subsequently $\sigma^2_{u0} = 0$)[5].

We can draw two conclusions from this. First, including the within-person means as a level 2 predictor in a multilevel autoregressive model is not logical, and therefore the results for contextual models presented above are less relevant in the current context. Second, the within-cluster slope will—without exception—differ from the between-cluster slope in multilevel autoregressive models: That is, while the between-cluster slope is (essentially) 1, the within-cluster slope is identical to the autocorrelation and will thus have to lie between -1 and 1 for a stationary process (Hamilton, 1994). This is illustrated in the right panel of **Figure 1**, which shows that the between-cluster slope is equal to 1, while the within-cluster slope (averaged across individuals) is smaller (in this case between 0 and 1).

Applying the reasoning offered by Raudenbush and Bryk (2002) and Enders and Tofighi (2007) about the effects of NC/GMC vs. CMC in case of standard multilevel models to

[4]Here it represents the expected difference in number of typos when comparing two participants who type the same number of words, while they differ one unit on the number of words they type per minute on average (meaning they have different levels of experience).

[5]In practice, the means on the outcome variable y_{ti} are virtually identical to the cluster means on the predictor $y_{t-1,i}$, as it is the same variable; slight difference may arise however, because the outcome runs over $t = 2, \ldots, T_i$ while the predictor runs over $t = 1, \ldots, T_i - 1$. As T becomes larger, these differences will become smaller.

the multilevel autoregressive model, one may thus be inclined to think that: (a) NC/GMC will result in an *overestimation* of the fixed effect (i.e., average) autoregressive parameter, since $\beta_W < \beta_B = 1$; and (b) CMC will remove the "contamination" of β_B, such that the fixed effect autoregressive parameter adequately represents β_W, that is, the (averaged or pooled) within-person autoregression ϕ. This would imply that CMC should be the preferred form of centering in a multilevel autoregressive model.

Note that we already discussed two other reasons for preferring CMC in case of the multilevel autoregressive model, that is, it allows us to model μ_i as a random effect, rather than the less meaningful $c_i = \mu_i(1 - \phi_i)$, and it allows us to include predictors for μ_i (rather than for c_i). Taken together, these seem very convincing reasons for preferring CMC over GMC/NC in a multilevel autoregressive model.

3. SIMULATIONS

We performed a series of simulations to investigate the effect of NC vs. CMC on the estimation of the within-cluster slope. We begin with the standard multilevel model to verify the claims made by Raudenbush and Bryk (2002) and Enders and Tofighi (2007), and to determine whether these also generalize to models with a random slope (as presented in Equations 24 and 25). Following this, we consider the effects of NC and CMC in the multilevel autoregressive model, both with a fixed and a random autoregressive parameter. In addition, we consider the effects of diverse factors, that is: sample sizes, the sign and strength of the autoregressive parameter, and a level 2 predictor for the autoregressive parameter. All our simulations were performed in R (R Development Core Team, 2009). To estimate the multilevel models, we used the function `lmer()` from the R-package lme4 (Bates and Sarkar, 2007).

3.1. SIMULATIONS FOR THE STANDARD MULTILEVEL MODEL

We begin with simulating data from the standard multilevel model with different within-cluster and between-cluster slopes, using Equation 19 for a model with a fixed within-cluster slope, and Equation 23 for a model with a random within-cluster slope. Our specific interest is in obtaining an appropriate estimate of the within-cluster slope, when this differs from the between-cluster slope. Hence, we want to verify that when the cluster means are not included as a level 2 predictor, the slope estimate obtained with NC is indeed a blend of the within-cluster and between-cluster slopes, while CMC (based on Equation 15) leads to a pure within-cluster slope estimate.

We used the following model parameter values: (a) the variance of the predictor x_{ti} within each cluster is 1, and the variance of the cluster means between the clusters is also 1; (b) the fixed effect within-cluster slope γ_{10} is 0.3; (c) the standard deviation of the within-cluster slope β_i is either 0 (i.e., fixed slope only model), or 0.1 (i.e., random slope model); (d) the between-cluster slope γ_{01} is 1; (e) the grand mean γ_{00} is zero; (f) the level 1 residual variance σ_e^2 was either 1 or 3; and (g) the level 2 residual variance for the intercept σ_{u0}^2 was either 1 or 0. The reason we considered 0 as well here, is because this would make the model more comparable to the multilevel autoregressive model we consider later on (see Equation 26). We set the number of clusters to 100, and the number of observations per cluster to 20.

The results are presented in **Table 2**: It includes the OLS estimate of the between-cluster slope (based on Equation 12), the OLS estimate of the within-cluster slope (based on Equation 13), and the fixed effects slope obtained with CMC and with NC. These confirm the point made by Raudenbush and Bryk (2002) and Enders and Tofighi (2007): While CMC leads to a slope estimate that is almost identical to the OLS within-cluster estimate and which adequately represents the actual within-cluster slope, the estimate obtained with NC is a blend of the within-cluster and the between-cluster slopes. Specifically, if the level 2 residual variance (i.e., σ_{u0}^2) becomes smaller relative to the level 1 residual variance (i.e., σ_e^2), the slope estimate is more strongly affected by the between-cluster slope. Furthermore, the results are very similar for models and data without a random slope (left part of **Table 2**), and with a random slope (right part of **Table 2**).

3.2. SIMULATIONS FOR THE MULTILEVEL AUTOREGRESSIVE MODEL

To determine whether the results reported above generalize to the multilevel autoregressive model, we considered the following scenarios. We simulated data using the model defined in Equations 1–4, with: (a) a fixed effects within-cluster slope of $\phi = 0.3$; (b) a standard deviation of the individual within-cluster slope ϕ_i of either 0 (for a model with a fixed autoregressive parameter only) or 0.1 (for a model with a random autoregressive parameter); (c) a level 2 variance of the intercept μ_i of 1, 3 or 9; and (d) a grand mean of 0. We used the same number of observations as in the previous simulations, that is, 100 clusters (i.e., persons here) and 20 observations per cluster (i.e., repeated measurements here). The results based on 1000 replications are presented in **Table 3**.

Table 2 | Estimates for fixed effect slope γ_{10}.

Estimation method	σ_{u0}^2	σ_e^2	Fixed slope only	Random slope
OLS between	1	1	0.963	0.965
	0	1	0.966	0.967
	0	3	0.965	0.966
OLS within	1	1	0.300	0.299
	0	1	0.298	0.300
	0	3	0.299	0.302
CMC (sample)	1	1	0.300	0.299
	0	1	0.298	0.300
	0	3	0.299	0.303
NC	1	1	0.323	0.323
	0	1	0.372	0.373
	0	3	0.492	0.489

Mean point estimates for fixed effects slope γ_{10} in a standard multilevel model with either a fixed slope only (left; $\beta_i = \gamma_{10}$), or with a random slope (right; $\beta_i = \gamma_{10} + u_{1i}$). True fixed effect within-cluster slope is $\gamma_{10} = 0.3$, and true between-cluster slope is $\gamma_{01} = 1$. Number of observations per cluster is 20; number of clusters is 100; number of replications is 1000. Estimation methods are: OLS between and within (Equations 12 and 13); centering per cluster (CMC) using the sample mean; and no centering (NC).

Table 3 | Estimates for fixed effect autoregressive parameter ϕ.

Estimation method	σ_e^2	Fixed slope only	Random slope
OLS within	1	0.230	0.233
	3	0.229	0.233
	9	0.228	0.233
CMC (sample)	1	0.231	0.229
	3	0.230	0.229
	9	0.229	0.229
NC	1	0.304	0.307
	3	0.304	0.306
	9	0.303	0.304

Mean point estimates for the fixed effects autoregressive parameter ϕ in a multilevel autoregressive model, with a fixed slope only (left; $\phi_i = \phi$), and with a random slope (right; $\phi_i = \phi + u_{1i}$). True fixed effect autoregressive parameter (i.e., the true within-cluster slope) is $\phi = 0.3$. Number of observations per person is 20; number of persons is 100; number of replications is 1000. Estimation methods are: OLS within (Equation 13); centering per cluster (CMC) using the sample mean; and no centering (NC).

As before, CMC leads to estimates that are very close to the OLS within-cluster estimates. However, for the multilevel autoregressive model, these are biased: That is, they underestimate the actual fixed effect autoregressive parameter (i.e., estimated bias between 0.069 and 0.071 for CMC). Surprisingly, NC leads to an estimate that is less biased (i.e., estimated bias between 0.003 and 0.007). In Appendix 2, this downward bias for the OLS within-cluster estimate in multilevel autoregressive model is confirmed analytically. Note further that whether or not inertia was random, did not affect the results substantially.

3.3. INVESTIGATING THE INFLUENCE OF OTHER FACTORS

To gain more insight in this matter, we considered three additional factors that may affect the estimation of the within-cluster slope in a multilevel autoregressive model. First, in addition to using the individual sample means in CMC (i.e., $\bar{y}_{\cdot i}$), we also considered the empirical Bayes estimator (also referred to as shrinkage estimator) of the individuals' means (i.e., $\hat{\mu}_i$, obtained with estimating the empty model first), and the true person means that were used to generate the data (i.e., μ_i; we considered this option here to see to what extent the results for CMC can be attributed to having to use an estimate of the individual's mean). Second, we considered different samples sizes, both with respect to number of persons N, and the number of repeated measures T. Third, we considered different strengths and signs of the fixed effects autoregressive parameter. Throughout we used the level 1 residual variance $\sigma_e^2 = 3$, the level 2 intercept variance $\sigma_{u0}^2 = 3$, and the level 2 slope variance $\sigma_{u1}^2 = 0.01$.

Based on the results presented in **Table 4**, we can conclude the following. First, CMC of the autoregressive predictor leads to bias, regardless of the kind of mean that is used (i.e., the sample estimate $\bar{y}_{\cdot i}$, the empirical Bayes estimate $\hat{\mu}_i$, or the true value μ_i). It is noteworthy that even using the true mean results in bias that is about the same as the bias obtained with the empirical Bayes estimate of the mean, while using the sample mean leads to only slightly more bias. In contrast, NC does not lead to (considerable) bias. Second, when using CMC, increasing the number of observations per person (i.e., T) leads to a decrease in bias, whereas the number of individuals N does not affect the bias. Third, the bias for CMC reported in **Table 4** is *always negative*, regardless of the actual value of ϕ, although the bias is largest when $\phi = 0.3$, and smallest when $\phi = -0.3$. This implies that in general, ϕ will be underestimated when CMC is used, and the bias is larger when ϕ is positive (which will often be the case in practice). This is also confirmed by the analytical results in Appendix 2.

With respect to the coverage rates of the 95% confidence intervals, we make the following two observations. First, while in general they are too low, for NC most coverage rates are above 0.900, while for all three forms of CMC they are much lower (which is not surprising, given the bias of CMC estimates). Second, while increasing T leads to higher coverage rates for the CMC approaches, increasing N actually leads to lower coverage rates. This result is explained by the fact that the standard errors decrease when N increases, while the bias remains unaffected by changes in N. Note that the pattern for the coverage rates obtained with NC is less clear.

3.4. INCLUDING A LEVEL 2 PREDICTOR OF THE AUTOREGRESSIVE PARAMETER

An important question when applying the multilevel autoregressive model is whether other variables predict individual differences in the autoregression (cf. Suls et al., 1998; Kuppens et al., 2010). Therefore, we performed an additional simulation study to determine the effect of CMC and NC on the estimation of the effect of a level 2 predictor on the autoregressive parameter.

Let z_i be a level 2 predictor that may have an effect on the individuals' average score μ_i, but more importantly, may have an effect on the individuals' autoregressive parameter ϕ_i. We assume this level 2 predictor is centered across people. When using NC, the model can be expressed as

$$\begin{aligned} y_{ti} &= c_i + \phi_i^n y_{t-1,i} + e_{ti} \\ c_i &= \gamma_{00}^n + \gamma_{01}^n z_i + u_{0i} \\ \phi_i^n &= \gamma_{10}^n + \gamma_{11}^n z_i + u_{1i} \end{aligned} \tag{27}$$

where γ_{00}^n is the overall intercept, and γ_{10}^n is the average autoregressive parameter (assuming the level 2 predictor z_i is centered). The regression coefficients γ_{01}^n and γ_{11}^n represent the effects of the level 2 predictor on the individuals' intercept c_i and their autoregressive parameter ϕ_i^n, respectively.

In contrast, when using CMC for the autoregressive predictor, the model can be defined as

$$\begin{aligned} y_{ti} &= \mu_i + \phi_i^c\left(y_{t-1,i} - \mu_i\right) + e_{ti} \\ \mu_i &= \gamma_{00}^c + \gamma_{01}^c z_i + u_{0i} \\ \phi_i^c &= \gamma_{10}^c + \gamma_{11}^c z_i + u_{1i} \end{aligned} \tag{28}$$

where γ_{00}^c now represents the grand mean, and γ_{10}^c is again the average autoregressive parameter (assuming the level 2 predictor

Table 4 | Bias and coverage rates for fixed autoregressive parameter ϕ in multilevel autoregressive model under diverse scenarios.

AR parameter	Sample size		Bias				$CR_{0.95}$			
	N	T	NC	C($\bar{y}_{.i}$)	C($\hat{\mu}_i$)	C(μ_i)	NC	C($\bar{y}_{.i}$)	C($\hat{\mu}_i$)	C(μ_i)
$\phi_i \sim N(0.3, 0.1)$	20	20	0.002	−0.072	−0.069	−0.068	0.928	0.762	0.785	0.787
		50	0.000	−0.027	−0.027	−0.026	0.940	0.900	0.901	0.898
		100	0.000	−0.013	−0.013	−0.013	0.932	0.932	0.932	0.932
	50	20	0.005	−0.071	−0.069	−0.067	0.893	0.480	0.512	0.518
		50	0.001	−0.027	−0.026	−0.026	0.936	0.800	0.804	0.805
		100	0.000	−0.013	−0.013	−0.013	0.946	0.902	0.902	0.903
	100	20	0.006	−0.070	−0.068	−0.066	0.892	0.196	0.227	0.242
		50	0.001	−0.027	−0.027	−0.027	0.930	0.623	0.630	0.637
		100	0.000	−0.013	−0.013	−0.013	0.930	0.851	0.854	0.851
$\phi_i \sim N(0, 0.1)$	20	20	0.001	−0.053	−0.050	−0.050	0.923	0.844	0.858	0.851
		50	−0.000	−0.020	−0.020	−0.020	0.944	0.912	0.915	0.911
		100	0.000	−0.010	−0.009	−0.009	0.929	0.926	0.926	0.925
	50	20	0.003	−0.052	−0.049	−0.049	0.922	0.700	0.727	0.725
		50	−0.001	−0.021	−0.021	−0.021	0.942	0.860	0.862	0.861
		100	0.000	−0.010	−0.010	−0.010	0.939	0.910	0.910	0.909
	100	20	0.003	−0.053	−0.051	−0.050	0.929	0.431	0.479	0.477
		50	0.000	−0.021	−0.021	−0.021	0.931	0.775	0.785	0.785
		100	0.000	−0.010	−0.010	−0.010	0.942	0.892	0.896	0.896
$\phi_i \sim N(-0.3, 0.1)$	20	20	0.003	−0.034	−0.031	−0.032	0.943	0.907	0.916	0.913
		50	0.000	−0.014	−0.013	−0.014	0.940	0.932	0.934	0.928
		100	0.001	−0.005	−0.005	−0.005	0.928	0.928	0.929	0.929
	50	20	0.000	−0.038	−0.035	−0.036	0.940	0.783	0.802	0.795
		50	0.000	−0.014	−0.014	−0.014	0.932	0.894	0.896	0.896
		100	0.000	−0.007	−0.006	−0.006	0.927	0.914	0.914	0.914
	100	20	0.000	−0.039	−0.036	−0.037	0.932	0.597	0.639	0.624
		50	0.000	−0.015	−0.015	−0.015	0.928	0.848	0.851	0.851
		100	0.000	−0.007	−0.007	−0.007	0.942	0.908	0.911	0.911

Bias and coverage rates of 95% confidence intervals ($CR_{0.95}$) based on 1000 replications. N refers to number of persons, T refers to number of observations per person. The random coefficient ϕ_i comes from a normal distribution, with mean ϕ (either 0.3, 0, or −0.3), and standard deviation 0.1. Results are obtained for: NC of the autoregressive predictor; C($\bar{y}_{.i}$) is CMC using the sample mean; C($\hat{\mu}_i$) is CMC using the empirical Bayes estimate; and C(μ_i) is CMC using the true mean per person (for comparison).

z_i is centered). The regression coefficients γ_{01}^c and γ_{11}^c represent the effects of the level 2 predictor on the individuals' means μ_i and their autoregressive parameters ϕ_i^c, respectively.

Based on the results from the previous simulations, we expect that CMC (as in Equation 28) will lead to a downward bias in the estimation of the average autoregressive parameter ϕ (i.e., γ_{10}^c will be an underestimate), while NC (as in Equation 27) is not associated with such bias (i.e., γ_{10}^n is an unbiased estimate of ϕ). However, the question here is how CMC and NC affect the estimation of the level 2 predictor on ϕ_i, that is, γ_{11}^c and γ_{11}^n.

We created a level 2 predictor with a mean of zero and a variance of 0.01. We chose this rather small variance for numerical reasons: Because the variance of ϕ_i is necessarily small (say about 0.01), having a level 2 predictor with a large variance may lead to numerical problems in estimating the regression coefficient γ_{11}. The mean autoregressive parameter ϕ was set to 0.3. The effect of the level 2 predictor z_i on the individual inertia parameters ϕ_i was set to 0.4. The other parameters were chosen such that the correlations between μ_i, ϕ_i and z_i were not unrealistically high (μ_i and ϕ_i were both correlated 0.37 with z_i, and 0.14 with each other). After generating μ_i and ϕ_i from z_i, we used Equations 1 and 3 to generate the data. The results for this simulation study are presented in **Table 5**.

The left part of the **Table 5** contains the results that reflect the bias. In line with our previous results, the average autoregressive parameter $\gamma_{10} = \phi$ is characterized by a downward bias when CMC is used for the autoregressive predictor, while NC leads to unbiased estimates. However, when considering the effect of CMC vs. NC on the estimation of γ_{11}, we see that CMC actually leads to *less bias* than NC. Note also that while increasing T reduces the bias obtained with NC, the effect of increasing N is not that clear (i.e., when $T = 20$, increasing N actually increases the bias).

The right part of **Table 5** contains the coverage rates of the 95% confidence intervals. As before, the coverage rates for the average autoregressive parameter obtained with CMC are lower than

Table 5 | Results for average autoregressive parameter ϕ and the effect of a level 2 predictor z_i on the autoregressive parameter ϕ_i.

N	T	Bias				$CR_{0.95}$			
		γ_{10}		γ_{11}		γ_{10}		γ_{11}	
		NC	CMC	NC	CMC	NC	CMC	NC	CMC
20	20	−0.007	−0.076	−0.050	−0.007	0.897	0.720	0.944	0.958
	50	−0.005	−0.030	−0.045	−0.026	0.922	0.856	0.939	0.944
	100	0.000	−0.013	−0.019	−0.008	0.923	0.909	0.942	0.944
50	20	0.004	−0.071	−0.068	−0.022	0.885	0.476	0.950	0.959
	50	0.001	−0.026	−0.032	−0.014	0.904	0.781	0.945	0.948
	100	0.000	−0.013	−0.018	−0.006	0.924	0.890	0.953	0.949
100	20	0.004	−0.071	−0.084	−0.036	0.918	0.170	0.921	0.940
	50	0.000	−0.027	−0.024	−0.003	0.907	0.628	0.944	0.955
	100	0.001	−0.012	−0.019	−0.008	0.928	0.832	0.939	0.942

Bias and coverage rate of 95% confidence intervals ($CR_{0.95}$) based on 1000 replications. Results for $\gamma_{10} = 0.3$, that is, the average autoregressive parameter ϕ, and for $\gamma_{11} = 0.4$, that is, the effect of a level 2 predictor on the autoregressive parameter, using NC and CMC (with sample mean) for the autoregressive predictor.

those obtained with NC. For γ_{11} the coverage rates obtained with NC are in general lower than those obtained with CMC (which was to be expected given the results for the bias).

3.5. CONCLUSION

The first set of simulations presented in this section clearly illustrated the point made by Raudenbush and Bryk (2002) and Enders and Tofighi (2007) in case of a standard multilevel model. In addition it was shown that the claims regarding the within-cluster slope generalize to the model with a random slope, in that CMC leads to an estimate of the within-cluster slope, whereas NC results in a blend of the within-cluster and the between-cluster slope. The second set of simulations was based on the multilevel autoregressive model and showed that while CMC still leads to results that are almost identical to the OLS-within estimate, both of these are biased with respect to the actual within-cluster slope (i.e., the autoregressive relationship).

Additional simulations showed that there is a downward bias regardless of the sign of ϕ, and that this bias is most severe when T is small, while N has little (if any) influence. This bias could not be attributed to the quality of the estimate of the individuals' means (as very similar results are obtained when using the true means μ_i for centering). Furthermore, these results were supported by the derived relationship between the OLS within-cluster slope estimate and the value of ϕ in Appendix 2. In contrast, NC does not lead to bias in the estimation of the autoregressive parameter, which implies that the obtained result is actually *not contaminated* by the between-cluster relationship, as is the case in regular multilevel analysis. Finally, when adding a level 2 predictor to the model, the results described above for the average autoregressive parameter remain intact, but for the effect of the level 2 predictor on the autoregressive parameter, NC actually results in bias, whereas CMC does not.

4. APPLICATION: PART 2

Returning to the empirical data that we introduced in the beginning of this paper, we are now able to study inertia in daily relationship specific PA and NA, and include a level 2 predictor for the individual differences in the means and the inertia. We used Relationship Satisfaction, which was obtained prior to the diary study, and standardized this level 2 predictor to facilitate interpretation (i.e., we subtracted the grand mean, and divided it by the grand standard deviation). We used the model based on CMC (see Equation 28), and summarized the results for all the fixed effects in **Table 6**. We also included the estimate of the fixed effects inertia obtained with NC in this table, as the simulations reported in this paper showed that this is an unbiased estimate of the average inertia, whereas the corresponding estimate obtained with CMC is negatively biased.

It shows that on average there is significant inertia in relationship specific PA and NA for both males and females (see γ_{10}^{n}). In addition, Relationship Satisfaction proved a significant positive predictor of mean levels of relationship specific PA in both males and females, and a significant negative predictor of mean levels of relationship specific NA in both males and females (see γ_{01}^{c}). Furthermore, Relationship Satisfaction is a significant negative predictor of inertia in relationship specific PA in males (but not in females), and in relationship specific NA in males and females (see γ_{11}^{c}). This implies that individuals who are less satisfied with their relationship, are characterized by more carryover of relationship specific NA, than individuals who are more satisfied with their relationship. In addition, males who are less satisfied with their relationship, are also characterized by more carryover in their relationship specific PA. While the latter may seem surprising at first—as it implies that elevated relationship specific PA tends to persist over time for males who are less satisfied with their relationship—it also implies that attenuated relationship specific PA tends to prevail, which could be considered undesirable. These results are in agreement with other findings regarding inertia reported by Koval et al. (2013), who found that the inertias of PA and NA are positively correlated, and Kuppens et al. (2012), who found that inertia of angry and dysphoric behavior, but also of happy behavior all predicted the onset of depression. Taken together, these results seem to confirm the

Table 6 | Results for multilevel autoregressive model with a level 2 predictor (with random effects).

		Males			Females		
		est.	**SD**	***t*-value**	**est.**	**SD**	***t*-value**
PA	γ_{00}^{c}	3.514	0.043	81.98	3.498	0.043	80.93
	γ_{01}^{c}	0.303	0.046	6.58	0.383	0.045	8.43
	γ_{10}^{c}	0.354	0.016	22.46	0.340	0.015	22.50
	γ_{10}^{n}	0.385	0.015	25.41	0.369	0.015	24.69
	γ_{11}^{c}	−0.045	0.017	−2.67	−0.015	0.016	−0.98
NA	γ_{00}^{c}	1.346	0.022	59.96	1.346	0.020	5.75
	γ_{01}^{c}	−0.072	0.024	−2.99	−0.132	0.022	−6.15
	γ_{10}^{c}	0.242	0.017	14.31	0.224	0.016	13.64
	γ_{10}^{n}	0.267	0.017	16.14	0.254	0.017	15.39
	γ_{11}^{c}	−0.046	0.017	−2.63	−0.060	0.016	−3.71

Parameter estimates, standard errors and t-values for the fixed effects parameters in a multilevel autoregressive model (with random intercept and slope), for males and females. Parameters include: (a) the grand mean (i.e., γ_{00}^{c}); (b) the effect of Relationship Satisfaction on the individuals' means (i.e., γ_{01}^{c}); (c) the average inertia obtained with CMC (i.e., γ_{10}^{c}), and with NC (i.e., γ_{10}^{n}); and (d) the effect of Relationship Satisfaction on the individuals' inertias (i.e., γ_{11}^{c}).

idea that inertia—whether in pleasant or unpleasant emotions—is a detrimental property of affect regulation, reflective of some maladaptive process.

5. DISCUSSION

Over the past two decades we have witnessed an exponential increase in the number of studies based on intensive longitudinal data in the social sciences. This development is triggered by the rapid development of electronic data collection methods based on hand-held computers, the internet, and—more recently—smart phones (Trull and Ebner-Priemer, 2013): As a result it has become relatively easy to gather large numbers of repeated measurements from a large sample of individuals. Such data differ from more traditional longitudinal data in two important ways: (1) intensive longitudinal data contain many more measurements per individual (i.e., often $T > 20$) than traditional longitudinal data (i.e., often $T < 10$); and (2) the measurements in intensive longitudinal data are typically spaced relatively close to each other in time (e.g., measurements are taken at a daily basis using a daily diary method, or even multiple times a day using experience method sampling), whereas traditional longitudinal data are characterized by much larger intervals between measurements (e.g., annual measurements are not uncommon). These differences reflect a different focus on part of the researchers: Whereas the purpose of many studies based on traditional longitudinal data is to discover broad underlying increasing or decreasing trends, the purpose of studies based on intensive longitudinal data is to gain more insight into the patterns of fluctuations in affect, behavior, and cognition in daily life (Bolger et al., 2003; Mehl and Conner, 2012).

One particular aspect of such patterns is referred to as inertia or autoregression, and represents the amount of carryover from one measurement occasion to the next (Suls et al., 1998). Diverse empirical studies have now shown that individual differences in inertia are meaningful with respect to the way people differ in their regulation of emotions and behavior. As the popular method for studying inertia is through a multilevel autoregressive model, an important research question in this area is whether the autoregressive predictor included at level 1 should be centered per person or not.

The current study shows through a series of simulations that CMC should be preferred if: (a) one wishes to obtain a meaningful intercept (i.e., an intercept that represents the individual's mean score over time, which can be interpreted as his/her trait score); and (b) the interest is in how the autoregressive parameter depends on a level 2 predictor. However, CMC should *not* be used when the interest is in whether or not there is an autoregressive relationship *on average* (i.e., across individuals).

In practice, researchers using a multilevel autoregressive model to study inertia are likely to be interested in various aspects of the model, including the individuals' means, the average autoregressive parameter, and the effect of a level 2 predictor on the individuals' means and autoregressive parameters. In that case, it may be wise to use both estimation procedures, as we did in the empirical application, and to use CMC for the estimation of the grand mean and the effect of the level 2 predictor on the individual means and autoregressive parameters, while NC results are used for determining whether there is an autoregressive effect on average. While this may be unconventional advice, it is based on the rather clear simulation results presented in this paper.

Given the recent interest in inertia, and its emerging recognition as a separate and valuable property of regulation that is related to but does not coincide with more traditionally studied process features such as the tendency to ruminate or the persistence of negative thoughts, we expect to see more work in this area. Hence, it is important to improve our ways to estimate average inertia, and individual differences therein. Specifically, the current study has shown that many of the inertia estimates reported in the literature may actually be underestimates of the true inertias, simply because the lagged autoregressive predictor was centered per person (e.g., Koval et al., 2012; Brose et al., 2014). Although this may not come as a surprise to those familiar with time series literature, as it has been known for a long time that estimates of autoregressive parameters are biased (cf., Orcutt, 1948; Marriott and Pope, 1954), it is an unexpected result from a multilevel perspective. Furthermore, it is of interest that the bias disappears when the lagged autoregressive predictor is not centered; in fact, this may be considered an important advantage of the multilevel approach over a two-step procedure in which during the first step individual time series models are estimated, while in the second step the individual parameters are combined into a population model.

Additional improvements in the study of inertia may come from taking measurement error into account—which is also likely to obscure the actual inertia of a process—and developing appropriate techniques for handling unequal intervals between the observations—which are a feature of certain intensive longitudinal data, and which may lead to less precise estimates when not taken into account, and therefor to more difficulty in detecting relationships between inertia and other person characteristics. When these issues are handled in an appropriate way, inertia may prove to be an even more important feature of regulatory processes in psychology than the existing studies already suggest.

Finally, note that the advice given here regarding CMC vs. NC or GMC *exclusively* applies to an autoregressive level 1 predictor: That is, if one includes other level 1 predictors, the common results based on Raudenbush and Bryk (2002) apply to them, meaning that CMC of these predictors should be preferred over NC or GMC if the within-cluster and between-cluster slopes are expected to differ, and one wants to obtain an estimate of the within-cluster slope.

ACKNOWLEDGMENT

This study was supported by the Netherlands Organization for Scientific Research (NWO; VIDI Grant 452-10-007). The authors thank Emilio Ferrer for kindly making his data available.

REFERENCES

Bates, D. and Sarkar, D. (2007). *lme4: Linear mixed-effects models using S4 classes. R package*. Available online at: http://CRAN.R-project.org/

Bolger, N., Davis, A., and Rafaeli, E. (2003). Diary methods: capturing life as it is lived. *Annu. Rev. Psychol.* 54, 579–616. doi: 10.1146/annurev.psych.54.101601.145030

Bolger, N. and Zuckerman, A. (1995). A framework for studying personality in the stress process. *J. Pers. Soc. Psychol.* 69, 890–902. doi: 10.1037/0022-3514.69.5.890

Brose, A., Schmiedek, F., and Kuppens, P. (2014). Emotional inertia contributes to depressive symptoms beyond perseverative thinking. *Cogn. Emot.* 13:1–12. doi: 10.1080/02699931.2014.916252. [Epub ahead of print].

Curran, P. J. and Bauer, D. J. (2011). The disaggregation of within-person and between-person effects in longitudinal models of change. *Annu. Rev. Psychol.* 62, 583–619. doi: 10.1146/annurev.psych.093008.100356

Enders, C. K. and Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychol. Methods* 12, 121–138. doi: 10.1037/1082-989X.12.2.121

Ferrer, E., Steele, J. S., and Hsieh, F. (2012). Analyzing the dynamics of affective dyadic interactions using patterns of intra- and interindividual variability. *Multivar. Behav. Res.* 47, 136–171. doi: 10.1080/00273171.2012.640605

Ferrer, E., and Widaman, K. (2008). "Dynamic factor analysis of dyadic affective processes with inter-group differences," in *Modeling Dyadic and Interdependent Data in the Developmental and Behavioral Sciences*, eds N. A. Card, J. P. Selig, and T. D. Little (New York, NY: Taylor & Francis Group), 107–137.

Gunthert, K. C., Cohen, L. H., Butler, A. C., and Beck, J. S. (2007). Depression and next-day spillover of negative mood and depressive cognitions following interpersonal stress. *Cogn. Ther. Res.* 31, 521–532. doi: 10.1007/s10608-006-9074-1

Hamaker, E. L. (2012). "Why researchers should think 'with in-person' a paradigmatic rationale," in *Handbook of Research Methods for Studying Daily Life*, eds M. R. Mehl and T. S. Conner (New York, NY: Guilford Publications), 43–61.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton, NJ: Princeton University Press.

Henquet, C., Van Os, J., Kuepper, R., Delespaul, P., Smits, M., à Campo, J., et al. (2010). Psychosis reactivity to cannabis use in daily life: an experience sampling study. *Br. J. Psychiatry* 196, 447–453. doi: 10.1192/bjp.bp.109.072249

Hox, J. J. (2010). *Multilevel Analysis: Techniques and Application, 2nd Edn.* New York, NJ: Taylor & Francis Group.

Kievit, R. A., Frankenhuis, W. E., Waldorp, L. J., and Borsboom, D. (2013). Simpson$\check{S}$s paradox in psychological science: a practical guide. *Front. Psychol.* 4:513. doi: 10.3389/fpsyg.2013.00513

Koval, P. and Kuppens, P. (2012). Changing emotion dynamics: individual differences in the effect of anticipatory social stress on emotional inertia. *Emotion* 12, 256–267. doi: 10.1037/a0024756

Koval, P., Kuppens, P., Allen, N. B., and Sheeber, L. (2012). Getting stuck in depression: the roles of rumination and emotional inertia. *Cogn. Emot.* 26, 1412–1427. doi: 10.1080/02699931.2012.667392

Koval, P., Pe, M., Meers, K., and Kuppens, P. (2013). Affective dynamics in relation to depressive symptoms: variable, unstable or inert? *Emotion* 13, 1132–1141. doi: 10.1037/a0033579

Kreft, I. G. G., de Leeuw, J., and Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivar. Behav. Res.* 30, 1–21. doi: 10.1207/s15327906mbr3001_1

Kuppens, P., Allen, N. B., and Sheeber, L. B. (2010). Emotional inertia and psychological maladjustment. *Psychol. Sci.* 21, 984–991. doi: 10.1177/0956797610372634

Kuppens, P., Sheeber, L. B., Yap, M. B. H., Whittle, S., Simmons, J., and Allen, N. B. (2012). Emotional inertia prospectively predicts the onset of depression in adolescence. *Emotion* 12, 283–289. doi: 10.1037/a0025046

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* 38, 963–974. doi: 10.2307/2529876

Larson, R. W. and Almeida, D. M. (1999). Emotional transmission in the daily lives of families: a new paradigm for studying family process. *J. Marriage Fam.* 61, 5–20. doi: 10.2307/353879

Marriott, F. H. C. and Pope, J. A. (1954). Bias in the estimation of autocorrelations. *Biometrika* 41, 390–402. doi: 10.1093/biomet/41.3-4.390

Mehl, M. R. and Conner, T. S., editors (2012). *Handbook of Research Methods for Studying Daily Life*. New York, NY: The Guilford Press.

Moberly, N. J. and Watkins, E. R. (2008). Ruminative self-focus and negative affect: An experience sampling study. *J. Abnorm. Psychol.* 117, 314–323. doi: 10.1037/0021-843X.117.2.314

Nezlek, J. B. (2001). Multilevel random coefficient analyses of event- and interval-contingent data in social and personality psychology research. *Pers. Soc. Psychol. Bull.* 27, 771–785. doi: 10.1177/0146167201277001

Orcutt, G. H. (1948). A study of the autoregressive nature of the time series used for Tinbergen's model of the economic system of the United States, 1919-1932. *J. R. Stat. Soc. Series B* 10, 1–53.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods, 2nd Edn.* Thousand Oaks, CA: Sage Publications.

Rovine, M. J., and Walls, T. A. (2006). "Multilevel autoregressive modeling of interindividual differences in the stability of a process," in *Models for Intensive Longitudinal Data*, eds T. A. Walls and J. L. Schafer (New York, NY: Oxford University Press), 124–147.

Singer, J. D. and Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York, NY: Oxford University Press. doi: 10.1093/acprof:oso/9780195152968.001.0001

Snijders, T. A. B. and Bosker, R. J. (2012). *Multilevel Analysis: An introduction to basic and advanced multilevel modeling, 2 Edn*. London: Sage Publishers.

Suls, J., Green, P., and Hillis, S. (1998). Emotional reactivity to everyday problems, affective inertia, and neuroticism. *Pers. Soc. Psychol. Bull.* 24, 127–136. doi: 10.1177/0146167298242002

Trull, T. J. and Ebner-Priemer, U. (2013). Ambulatory assessment. *Annu. Rev. Clin. Psychol.* 9, 151–176. doi: 10.1146/annurev-clinpsy-050212-185510

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York, NY: Springer-Verlag.

Wang, L., Hamaker, E. L., and Bergman, C. S. (2012). Investigating inter-individual difference in short-term intra-individual variability. *Psychol. Methods* 17, 567–581. doi: 10.1037/a0029317

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 07 October 2014; accepted: 03 December 2014; published online: 06 January 2015.

*Citation: Hamaker EL and Grasman RPPP (2015) To center or not to center? Investigating inertia with a multilevel autoregressive model. Front. Psychol. **5**:1492. doi: 10.3389/fpsyg.2014.01492*

This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology.

Copyright © 2015 Hamaker and Grasman. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

APPENDIX 1

To show that the model expressed in Equations 5, 6 and 7, is structurally equivalent to the model in Equations 1–4, we first make use of the fact that $y_{t-1,i} - \mu_i = a_{t-1,i}$, such that we can rewrite Equation 3 as

$$\begin{aligned} a_{ti} &= \phi_i a_{t-1,i} + e_{ti} \\ &= \phi_i(y_{t-1,i} - \mu_i) + e_{ti} \\ &= \phi_i y_{t-1,i} - \phi_i \mu_i + e_{ti}. \end{aligned} \tag{A1}$$

Entering this in Equation 1, we can write

$$\begin{aligned} y_{ti} &= \mu_i - \phi_i \mu_i + \phi_i y_{t-1,i} + e_{ti} \\ &= (1-\phi_i)\mu_i + \phi_i y_{t-1,i} + e_{ti}, \end{aligned} \tag{A2}$$

which shows that

$$c_i = (1-\phi_i)\mu_i. \tag{A3}$$

This is a standard result from the time series literature on the first-order autoregressive model (cf. Hamilton, 1994).

APPENDIX 2

The rather unexpected results regarding CMC (i.e., centering the autoregressive predictor per person actually leads to bias in estimating the autoregressive parameter), was confirmed by the near identical results when using OLS. Below we show that OLS indeed leads to bias in the estimation of the autoregressive parameter ϕ. For simplicity of the presentation we will not make notational distinction between a random variable and it's (observed) value here.

Note that the OLS estimate of the regression model $y_i = b_0 + b_1 x_i + e_i$ can be expressed as $\hat{b}_1 = \text{cov}(x_i, y_i)/\text{var}(x_i)$. In a similar fashion, the OLS estimate of the within-person relationship ϕ in the autoregressive multilevel model can be expressed as the covariance between the person-centered predictor variable y_{ti} and the person-centered outcome variable $y_{i,t+1}$, divided by the variance of the person-centered predictor variable. To this end, let $T_i^* = T_i - 1$, and let

$$\bar{y}_{i\cdot(1)} = \frac{1}{T_i^*}\sum_{t=1}^{T_i^*} y_{ti} \quad \text{and} \quad \bar{y}_{i\cdot(2)} = \frac{1}{T_i^*}\sum_{t=1}^{T_i^*} y_{i,t+1} \tag{A4}$$

represent the estimated person means of the predictor variable y_{ti} and the outcome variable $y_{i,t+1}$, respectively. Then the OLS estimator of ϕ can be expressed as

$$\begin{aligned} \hat{\phi} &= \frac{\sum_{i=1}^n \left\{ \frac{1}{T_i^*} \sum_{t=1}^{T_i^*} \left(y_{ti} - \bar{y}_{i\cdot(1)} \right) \left(y_{i,t+1} - \bar{y}_{i\cdot(2)} \right) \right\}}{\sum_{i=1}^n \left\{ \frac{1}{T_i^*} \sum_{t=1}^{T_i} \left(y_{ti} - \bar{y}_{i\cdot(1)} \right)^2 \right\}} \\ &= \frac{\frac{1}{n}\sum_{i=1}^n \left\{ \frac{1}{T_i^*} \sum_{t=1}^{T_i - 1} \left(y_{ti} y_{it+1} - T_i^* \bar{y}_{i\cdot(1)} \bar{y}_{i\cdot(2)} \right) \right\}}{\frac{1}{n}\sum_{i=1}^n \left\{ \frac{1}{T_i^*} \sum_{t=1}^{T_i} \left(y_{ti}^2 - T_i^* \bar{y}_{i\cdot(1)}^2 \right) \right\}}, \end{aligned} \tag{A5}$$

To derive the asymptotic bias of this estimator, we begin by deriving the numerator of Equation A5. To this end, we first consider the conventional estimate of the covariance between y_{ti} and $y_{i,t+1}$ per person, that is,

$$s_i(t+1, t) = \frac{1}{T_i^*}\sum_{t=1}^{T_i^*} \left\{ y_{ti} y_{it+1} - T_i^* \bar{y}_{i\cdot(1)} \bar{y}_{i\cdot(2)} \right\}.$$

Taking the expectation of this covariance, conditional on i, gives

$$\begin{aligned} E\left[s_i(t+1, t) |\, i \right] = \frac{1}{T_i^*}\sum_{t=1}^{T_i^*} \Big\{ & E\left[y_{ti} y_{i,t+1} |\, i \right] \\ & - T_i^* E\left[\bar{y}_{i\cdot(1)} \bar{y}_{i\cdot(2)} |\, i \right] \Big\}. \end{aligned} \tag{A6}$$

Focussing on the first expectation on the right-hand side of Equation A6, we make use of the fact that $y_{i,t+1} = c_i + \phi_i y_{ti} + e_{i,t+1}$, such that we can write

$$\begin{aligned} E\left[y_{ti} y_{i,t+1} | i \right] &= E\left[y_{ti} \left\{ c_i + \phi_i y_{ti} + e_{i,t+1} \right\} | i \right] \\ &= E\left[y_{ti} c_i | i \right] + E\left[\phi_i y_{ti}^2 | i \right] + E\left[y_{ti} e_{i,t+1} | i \right] \\ &= (1-\phi_i)\mu_i E\left[y_{ti} | i \right] + \phi_i E\left[y_{ti}^2 | i \right] \\ &= (1-\phi_i)\mu_i^2 + \phi_i(\sigma_i^2 + \mu_i^2) \\ &= \mu_i^2 + \phi_i \sigma_i^2, \end{aligned} \tag{A7}$$

where $c_i = \mu_i(1-\phi_i)$, $\mu_i = E(y_{ti}|i)$ and $\sigma_i^2 = \text{Var}(y_{ti}|i) \propto \frac{1}{1-\phi_i^2}$.

The second expectation on the right-hand side of Equation A6 can be rewritten (using the geometric series), to obtain

$$\begin{aligned} E\left[\bar{y}_{i\cdot(1)} \bar{y}_{i\cdot(2)} |\, i \right] &= \frac{1}{T_i^* T_i^*} \sum_{t=1}^{T_i^*} \sum_{\tau=1}^{T_i^*} E\left[y_{ti} y_{i,\tau+1} |\, i \right] \\ &= \frac{1}{T_i^* T_i^*} \sum_{t=1}^{T_i^*} \sum_{u=1}^{T_i^*} \left\{ \phi_i^{|t-u-1|} \sigma_i^2 + \mu_i^2 \right\} \\ &= \sigma_i^2 \, \frac{T_i^*(1-\phi_i^2) - (1+\phi_i^2)(1-\phi_i^{T_i^*})}{T_i^{*2}(1-\phi_i)^2} \\ &\quad + \mu_i^2. \end{aligned} \tag{A8}$$

Inserting the expression in Equations A7 and A8 in Equation A6, the expected value for the covariance conditional on person i can be expressed as

$$E\left[s_i(t+1, t) |\, i \right] = \sigma_i^2 \left\{ \phi_i - \frac{T_i^*(1-\phi_i^2) - (1+\phi_i^2)(1-\phi_i^{T_i^*})}{T_i^{*2}(1-\phi_i)^2} \right\} \tag{A9}$$

In a similar way, the expected value of the variance conditional on person i can be obtained, resulting in

$$E\big[s_i(t,t)|\ i\big] = \sigma_i^2\ \left\{1 - \frac{T_i(1-\phi_i^2) - 2\phi_i(1-\phi_i^{T_i})}{T_i^2(1-\phi_i)^2}\right\}. \quad \text{(A10)}$$

By the law of large numbers, as the number of participants $n \to \infty$, the numerator on the right-hand side of Equation A5 converges in probability to $E[s_i(t+1,t)] = E\{E[s_i(t+1,t)|\ i]\}$, while the denominator converges in probability to $E[s_i(t,t)] = E\{E[s_i(t,t)|\ i]\}$. Therefore,

$$\hat{\phi} \xrightarrow{p} \frac{E\big[s_i(t+1,t)\big]}{E\big[s_i(t,t)\big]} = \frac{E\left[\sigma_i^2\left\{\phi_i - \frac{T_i^*(1-\phi_i^2) - (1+\phi_i^2)(1-\phi_i^{T_i^*})}{T_i^{*2}(1-\phi_i)^2}\right\}\right]}{E\left[\sigma_i^2\left\{1 - \frac{T_i(1-\phi_i^2) - 2\phi_i(1-\phi_i^{T_i})}{T_i^2(1-\phi_i)^2}\right\}\right]}. \quad \text{(A11)}$$

To show that in general the asymptotic bias will be negative, for simplicity, we assume T_i is large enough to treat $T_i^* \approx T_i$. Then we have to show that

$$\frac{E\left[\sigma_i^2\left\{\phi_i - \frac{T_i^*(1-\phi_i^2) - (1+\phi_i^2)(1-\phi_i^{T_i^*})}{T_i^{*2}(1-\phi_i)^2}\right\}\right]}{E\left[\sigma_i^2\left\{1 - \frac{T_i(1-\phi_i^2) - 2\phi_i(1-\phi_i^{T_i})}{T_i^2(1-\phi_i)^2}\right\}\right]} \leq \phi.$$

As the denominator is always positive[6] when $T_i \geq 2$, this is equivalent to showing that

$$E\Big[\sigma_i^2(\phi_i - \phi)\Big] + E\left[\sigma_i^2 \frac{T_i(1-\phi)(1-\phi_i^2) - (1-2\phi\phi_i+\phi_i^2)(1-\phi_i^{T_i})}{T_i^2(1-\phi_i)^2}\right] \geq 0.$$

The first term on the left-hand side is the sum of the covariance between autoregressive parameter ϕ_i and the variance of the series σ_i^2. Note that $\sigma_i^2 = \sigma_e^2/(1-\phi_i^2)$, which implies that σ_i^2 and ϕ_i are correlated. For symmetric distributions of ϕ_i around ϕ, using a Taylor expansion of $\frac{\phi_i - \phi}{1-\phi_i^2}$, we can show[7] that the correlation is positive if $\phi > 0$ and negative if $\phi < 0$. Therefore, assuming a density $f_\Phi(\phi_i)$ that is (approximately) symmetric about ϕ, the first term should be deemed positive if $\phi > 0$, and negative, if $\phi < 0$.

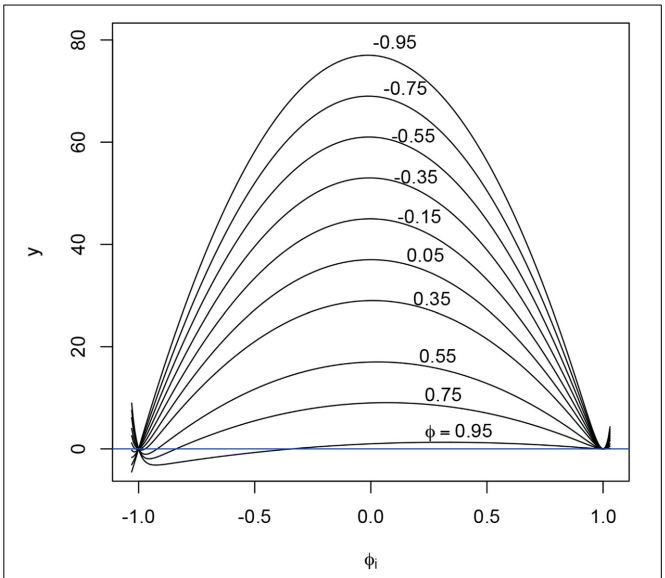


FIGURE A1 | Numerator of expectation, [that is, $y = T(1-\phi)(1-\phi_i^2) - (1-2\phi\phi_i+\phi_i^2)(1-\phi_i^T)$] plotted against ϕ_i, for $T = 40$ and different values of ϕ (i.e., average ϕ_i). Note that only for $\phi = 0.95$ the numerator becomes negative on a substantial portion of the interval $(-1, 1)$. See text for implications.

Regarding the second term, since the denominator in this term is always positive, the only way this expected value can be negative is if on a substantial portion of the support of the density $f_\Phi(\phi)$ of ϕ_i the numerator is negative. Since no closed form expression for this region can be found in terms of T and ϕ, below we plot the numerator for $T = 40$ and different values of ϕ in **Figure A1**. The interval for ϕ is limited to $(-1, 1)$ by the requirement of stationarity. The picture is slightly different for uneven T, but since the term $1-\phi^T$ only really matters near the edges of the interval, it has little effect on the global shape. It is clear that only for extreme values of ϕ ($\phi > 0.9$) a substantial portion of the numerator is negative, but then also only at the lower end of the interval. This means that for any reasonable $f_\Phi(\phi)$ (which incidentally must have $\int_{-1}^{1} \phi f_\Phi(\phi) d\phi = E(\phi_i) > 0.9$ and therefore cannot have a large probability mass in the area where the numerator is negative), the numerator will be positive and hence the second term will be positive. As T_i grows larger, the always positive term $T_i(1-\phi)(1-\phi_i^2)$ in the numerator becomes much larger than the term $(1-2\phi\phi+\phi^2)(1-\phi^T)$. Hence for values of T_i larger than 40, the negative region of the numerator in the expectation vanishes. As a result, the estimator $\hat{\phi}$ will in most cases underestimate the real value of ϕ.

[6]This follows if $[T(1-\phi^2) - 2(1-\phi^T)]/[T^2(1-\phi)^2] < 1 \iff T(1-\phi^2) - 2(1-\phi^T) < T^2(1-\phi)^2$, or $0 < T(T-1) - 2(T^2-1)\phi + T(T+1)\phi^2 - 2\phi^{T+1}$ for all ϕ. The latter may be seen to be true by graphing the function, or by differentiation. To this end, let's denote the right-hand side by Q. The second derivative is $Q'' = 2T(T+1)(1-\phi^{T-1})$ which is clearly always positive when $\phi \in (-1, 1)$. Hence, the first order derivative $Q' = -2(T^2-1) + 2T(T+1)\phi - 2(T+1)\phi^T$ is increasing on $(-1, 1)$ and reaches a maximum of 0 at $\phi = 1$. At any ϕ lower than $\phi = 1$ therefore, $y' < 0$ (for instance at $\phi = 0$, $Q' = -2(T^2-1) < 0$ assuming $T > 1$) which implies that Q must be strictly decreasing on $(-1, 1)$. In fact, Q reaches a minimum at $Q' = 0 \iff \phi = 1$, at which point $Q = 0$. Therefore, the minimum of the right hand side of the inequality is always positive, and so the inequality holds for all $\phi \in (-1, 1)$, as required.

[7]The Maclaurin series of $\frac{\phi_i - \phi}{1-\phi_i^2} = \frac{u_{1i}}{1-(\phi+u_{1i})^2} = \sum_{k=0}[(1-\phi)^{-k} - (-1-\phi)^{-k}]u_{1i}^k$, hence $E\{\frac{u_{1i}}{1-(\phi+u_{1i})^2}\} = \sum_{k=0}[(1-\phi)^{-k} - (-1-\phi)^{-k}]E\{u_{1i}^k\}$. For u_{1i} symmetrically distributed about 0, the odd moments are zero and the even moments are positive, and so the sign of $E\{\frac{u_{1i}}{1-(\phi+u_{1i})^2}\}$ only depends on the sign of $(1-\phi)^{-k} - (-1-\phi)^{-k}$ where $k = 2m$ even. It is readily verified that this is negative if and only if $\phi < 0$.

From Equation A11 it can be seen that as the lengths T_i of the observed series increase without bound, $\hat{\phi}$ converges in probability to $E\{(\sigma_i^2 + \mu_i^2)\phi_i\}/E\{(\sigma_i^2 + \mu_i^2)\}$. The disconcerting consequence is that the OLS estimator may be biased, even if an infinity number of samples is obtained!

THE CASE OF $\sigma_{u1}^2 = 0$

In the first set of simulations for the multilevel autoregressive model, $\sigma_{u1}^2 = 0$—that is, all individuals were characterized by the same autoregressive parameter (i.e., $\phi_i = \phi$ with probability 1). In this case, setting $T_i = T$ without loss of generality, the above inequality simplifies to

$$E\left[\sigma_i^2 \frac{T(1-\phi)(1-\phi^2)-(1-\phi^2)(1-\phi^T)}{T^2(1-\phi)^2}\right]$$
$$= \left(1 - \phi^2\right)\frac{T(1-\phi)-(1-\phi^T)}{T^2(1-\phi)^2} E[\mu_i^2] \geq 0,$$

where the first expectation dropped out because $E[\sigma_i^2(\phi_i - \phi)] = 0$ (since $\phi_i - \phi = 0$ when $\phi_i = \phi$).

The above inequality is satisfied if $g = T(1-\phi) - (1-\phi^T) \geq 0$. This is true for all $-1 \leq \phi \leq 1$, as then $g' = -T + T\phi^{T-1} \leq 0$ for all $T = 1, 2, \ldots$, and g achieves a minimum of 0 at $\phi = 1$. Hence, in this case, the estimator of ϕ is always biased downwards.

frontiers in Psychology

ORIGINAL RESEARCH
published: 28 July 2015
doi: 10.3389/fpsyg.2015.01038

Incorporating measurement error in $n = 1$ psychological autoregressive modeling

Noémi K. Schuurman[1]*, *Jan H. Houtveen*[2] and *Ellen L. Hamaker*[1]

[1] Department of Methodology and Statistics, Utrecht University, Utrecht, Netherlands, [2] Academic Centre of Psychiatry, Groningen University, Groningen, Netherlands

Measurement error is omnipresent in psychological data. However, the vast majority of applications of autoregressive time series analyses in psychology do not take measurement error into account. Disregarding measurement error when it is present in the data results in a bias of the autoregressive parameters. We discuss two models that take measurement error into account: An autoregressive model with a white noise term (AR+WN), and an autoregressive moving average (ARMA) model. In a simulation study we compare the parameter recovery performance of these models, and compare this performance for both a Bayesian and frequentist approach. We find that overall, the AR+WN model performs better. Furthermore, we find that for realistic (i.e., small) sample sizes, psychological research would benefit from a Bayesian approach in fitting these models. Finally, we illustrate the effect of disregarding measurement error in an AR(1) model by means of an empirical application on mood data in women. We find that, depending on the person, approximately 30–50% of the total variance was due to measurement error, and that disregarding this measurement error results in a substantial underestimation of the autoregressive parameters.

Keywords: autoregressive modeling, $n = 1$, measurement error, Bayesian modeling, idiographic, time series analysis

OPEN ACCESS

Edited by:
Craig Speelman,
Edith Cowan University, Australia

Reviewed by:
Emanuele Olivetti,
Bruno Kessler Foundation, Italy
James Stamey,
Baylor University, USA

***Correspondence:**
Noémi K. Schuurman,
Methodology and Statistics, Utrecht University, PO Box 80140, 3508 TC Utrecht, Netherlands
n.k.schuurman@uu.nl

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *18 May 2015*
Accepted: *07 July 2015*
Published: *28 July 2015*

Citation:
Schuurman NK, Houtveen JH and Hamaker EL (2015) Incorporating measurement error in n = 1 psychological autoregressive modeling. Front. Psychol. 6:1038. doi: 10.3389/fpsyg.2015.01038

1. Introduction

The dynamic modeling of processes at the within-person level is becoming more and more popular in psychology. The reason for this seems to be the realization that inter-individual differences, in many cases, are not equal to intra-individual differences. Indeed, studies that compare interindividual differences and intraindividual differences usually do not harbor the same results, exemplifying that conclusions based on studies of group averages (including cross-sectional studies and panel data studies), cannot simply be generalized to individuals (Nezlek and Gable, 2001; Borsboom et al., 2003; Molenaar, 2004; Rovine and Walls, 2006; Kievit et al., 2011; Madhyastha et al., 2011; Ferrer et al., 2012; Hamaker, 2012; Wang et al., 2012; Adolf et al., 2015).

The increased interest in analyses at the within-person level, and the increasing availability of technology for collecting these data, has resulted in an increase in psychological studies that collect intensive longitudinal data, consisting of many (say 25 or more) repeated measures from one or more individuals. A popular way to analyze these data currently is by autoregressive time series (AR) modeling, either by modeling the repeated measures for a single individual using classical $n = 1$ AR models, or by using multilevel extensions of these models, with the repeated measures for each individual modeled at level 1, and individual differences modeled at level 2

(Cohn and Tronick, 1989; Suls et al., 1998; Nezlek and Gable, 2001; Nezlek and Allen, 2006; Rovine and Walls, 2006; Moberly and Watkins, 2008; Kuppens et al., 2010; Lodewyckx et al., 2011; Madhyastha et al., 2011; Wang et al., 2012; De Haan-Rietdijk et al., 2014). In an AR model of order 1 [i.e., an AR(1) model], a variable is regressed on a lagged version of itself, such that the regression parameter reflects the association between this variable and itself at the previous measurement occasion (c.f., Hamilton, 1994; Chatfield, 2004). The reason for the popularity of this model may be the natural interpretation of the resulting AR parameter as inertia, that is, resistance to change (Suls et al., 1998). Resistance to change is a concept which is considered to be relevant to many psychological constructs and processes, including attention, mood and the development of mood disorders, and the revision of impressions and opinions (Geller and Pitz, 1968; Goodwin, 1971; Suls et al., 1998; Kirkham et al., 2003; Kuppens et al., 2010; Koval et al., 2012).

However, a problem with the regular AR(1) model is that it does not account for any measurement errors present in the data. Although AR models incorporate residuals, which are referred to as "innovations" or "dynamic errors," these residuals are to be distinguished from measurement error. Simply put, the distinction between dynamic errors and measurement errors is that dynamic errors carry over to next measurement occasions through the autoregressive relationship, while measurement errors are specific to one measurement occasion. Therefore, even though taking measurement errors into account is considered business as usual in many psychological studies of interindividual differences, it is largely neglected in AR modeling. Two exceptions are formed by Wagenmakers (2004) and Gilden (2001)[1], both of which concern studies on reaction time and accuracy in series of cognitive tasks. Gilden notes that there is evidence that some variance in reaction time is random (measurement) error as a result of key-pressing in computer tasks. Measurement error however is not limited to "accidentally" pressing the wrong button or crossing the wrong answer, but is made up of the sum of all the influences of unobserved factors on the current observation, that do not carry-over to the next measurement occasion. Disregarding measurement error distorts the estimation of the effects of interest (Staudenmayer and Buonaccorsi, 2005). This is quite problematic, considering that in psychological studies it is often impossible to directly observe the variable of interest, and it therefore seems likely (and this seems generally accepted among psychological researchers) that psychological research in general is prone to having noisy data.

The aim of this study is therefore three-fold. First, we aim to emphasize the importance of considering measurement error in addition to dynamic error in intensive longitudinal studies, and illustrate the effects of disregarding it in the case of the $n = 1$ autoregressive model. Second, we aim to compare two modeling strategies for incorporating measurement errors: (1) fitting an autoregressive model that includes a white noise term (AR+WN), and (2) fitting an autoregressive moving average (ARMA) model. These modeling strategies are the two most frequently suggested in the literature (e.g., in mathematical statistics, control engineering, and econometrics, c.f., Granger and Morris, 1976; Deistler, 1986; Chanda, 1996; Swamy et al., 2003; Staudenmayer and Buonaccorsi, 2005; Chong et al., 2006; Costa and Alpuim, 2010; Patriota et al., 2010). Third, our aim is to compare the performance of these models for a frequentist and a Bayesian estimation procedure. Specifically, for the frequentist procedure we will focus on a Maximum Likelihood (ML) procedure based on the state-space modeling framework, which is a convenient modeling framework for psychological longitudinal modeling, as it readily deals with missing data, and is easily extended to multivariate settings, or to include latent variables (Harvey, 1989). The Bayesian alternative shares these qualities, and has the additional advantage that the performance of the estimation procedure is not dependent on large samples (Dunson, 2001; Lee and Wagenmakers, 2005), while the performance of the frequentist ML procedure depends on asymptotic approximations, and in general requires large samples. This is convenient for the modeling of intensive longitudinal data, given that large amounts of repeated measures are often difficult to obtain in psychological studies. By means of a simulation study we will evaluate the parameter recovery performance of the Bayesian procedure for the ARMA(1,1) and the AR+WN model, and compare it to the ML procedure.

This paper is organized as follows. We start by introducing the AR(1) model, ARMA(1,1) model, and the AR(1)+WN model, and discussing their connections. After that, we present the methods for the simulation study, followed by the results. We present an empirical application concerning the daily mood of eight women, in order to further illustrate the consequences of disregarding measurement error in practice, and we end with a discussion.

2. Models

In this section we present the AR(1) model, and explain the difference between the dynamic errors that are incorporated in the AR(1) model, and measurement errors. After that we will introduce models that incorporate measurement errors, namely the autoregressive model with an added white noise term (AR(1)+WN model), and the autoregressive moving average (ARMA) model.

2.1. The AR(1) Model

In order to fit an AR model, a large number of repeated measures is taken from one individual. Each observation, or score, y_t in the AR model consists of a stable trait part—the mean of the process denoted as μ, and a state part $\tilde{y}_t$ that reflects the divergence from that mean at each occasion. In an AR model of order 1, the state of the individual at a specific occasion $\tilde{y}_t$ depends on the previous state $\tilde{y}_{t-1}$, and this dependency is modeled with the AR parameter ϕ. Specifically, the AR(1) model can be specified as

$$\begin{aligned} y_t &= \mu + \tilde{y}_t \\ \tilde{y}_t &= \phi \tilde{y}_{t-1} + \epsilon_t \end{aligned} \quad (1)$$

[1]Other exceptions are of course dynamic factor models, and other latent variable models in which the measurement structure for multiple items is explicitly modeled. Here we focus on applications in which each construct is measured with one variable.

$$\epsilon_t \sim N\left(0, \sigma_\epsilon^2\right). \quad (2)$$

For a graphical representation of the model, see **Figure 1A**. A positive value for ϕ indicates that the score at the current occasion will be similar to that at the previous occasion— and the higher the positive value for ϕ, the more similar the scores will be. Therefore, a positive AR parameter reflects the inertia, or resistance to change, of a process (Suls et al., 1998). A positive AR parameter could be expected for many psychological processes, such as that of mood, attitudes, and (symptoms of) psychological disorders. A negative ϕ indicates that if an individual has a high score at one occasion, the score at the next occasion is likely to be low, and vice versa. A negative AR parameter may be expected for instance in processes that concern intake, such as drinking alcoholic beverages: If an individual drinks a lot at one occasion, that person may be more likely to cut back on alcohol the next occasion, and the following occasion drink a lot again, and so on Rovine and Walls (2006). An AR parameter close to zero indicates that a score on the previous occasion does not predict the score on the next occasion. Throughout this paper we consider stationary models, which implies that the mean and variance of y are stable over time, and ϕ lies in the range from -1 to 1 (Hamilton, 1994). The innovations ϵ_t reflect that component of each state score $\tilde{y}_t$ that is unpredictable from the previous observation. The innovations ϵ_t are assumed to be normally distributed with a mean of zero and variance σ_ϵ^2.

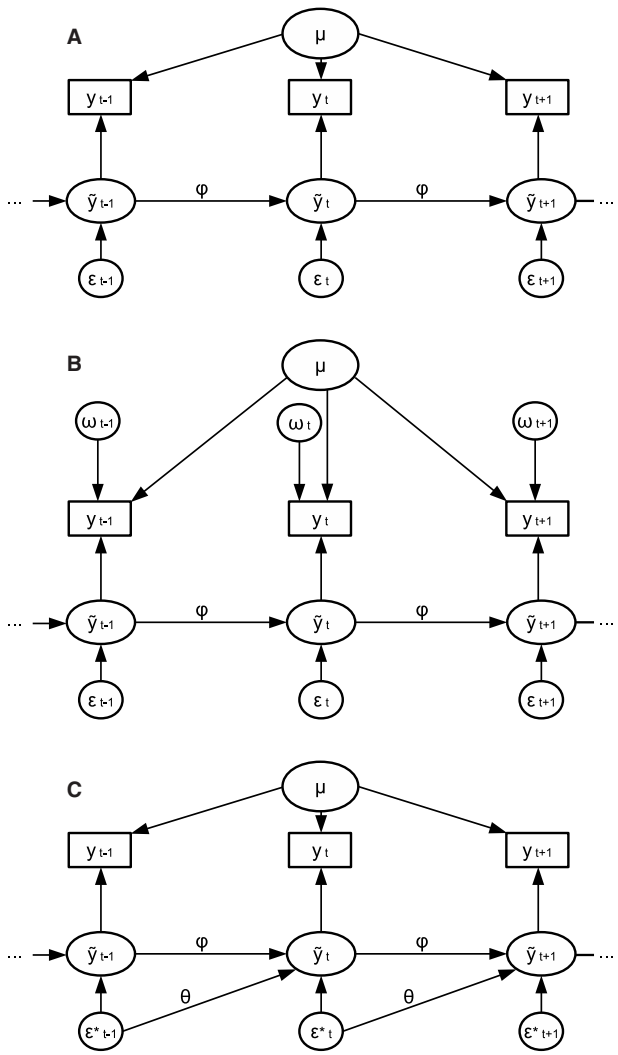


FIGURE 1 | (A) Graphical representation of an AR(1) model. **(B)** Graphical representation of an AR(1)+WN model. **(C)** Graphical representation of an ARMA(1,1) model.

2.2. Dynamic Errors vs. Measurement Errors

The innovations ϵ_t perturb the system and change its course over time. Each innovations is the result of all unobserved events that impact the variable of interest at the current measurement occasion, of which the impact is carried over through the AR effect to the next few measurement occasions. Take for example hourly measurements of concentration: Unobserved events such as eating a healthy breakfast, a good night sleep the previous night, or a pleasant commute, may impact concentration in the morning, resulting in a heightened concentrating at that measurement occasion. This heightened concentration may then linger for the next few measurement occasions as a result of an AR effect. In other words, the innovations ϵ_t are "passed along" to future time points via ϕ, as can be seen from **Figure 1A**, and this is why they are also referred to as "dynamic errors."

Measurement errors, on the other hand, do not carry over to next measurement occasions, and their effects are therefore restricted to a single time point. This can also be seen from **Figure 1B**: The dynamic errors are passed from y_{t-1} to y_t through the AR effect while the measurement errors ω_t are specific to each observation. Classical examples of measurement error, which are moment-specific, are making an error while filling in a questionnaire, or accidentally pressing a (wrong) button during an experiment (e.g., Gilden, 2001). However, any unobserved effect of which the influence is not carried over to the next measurement occasion may also be considered as measurement error, rather than dynamic error. The only distinguishing characteristic of measurement errors and dynamic errors is that the latter's influence lingers for multiple measurement occasions. Therefore, in practice, what unobserved effects will end up as measurement error, and what effects will end up as dynamic error, will depend largely on the measurement design of the study, such as on the frequency of the repeated measures that are taken. For example, some unobserved effects may carry-over from minute to minute (e.g., having a snack, listening to a song), but not from day to day—if measurements are then taken every minute, these unobserved effects will end up in the dynamic error term, but if measurements are taken daily, such effects will end up in the measurement error term. As such, the more infrequent measurements are taken, the more measurement errors one can expect to be present in the data, relative to the dynamic errors.

In psychological research measurement is complicated, and likely to be noisy. As such, the contribution of measurement error variance to the total variance of the measured process may be considerable. Ignoring this contribution will result in biased parameter estimates. Staudenmayer and Buonaccorsi (2005) have

shown that in the case of an AR(1) model, ϕ will be biased toward zero. Specifically, the estimated AR coefficient $\hat{\phi}$ will be equal to $(1 - \lambda) * \phi$, where ϕ is the true AR parameter and λ is the proportion of measurement error variance to the total variance. Hence, in order to prevent the measurement error from biasing estimates of ϕ, it is necessary to take measurement error into account in the modeling procedure. This approach has two advantages: First, it leads to less biased estimates of ϕ, and second, it allows us to investigate to what extent the measurements are determined by measurement error.

2.3. Incorporating Measurement Error: The AR(1)+WN Model

A relatively simple way to incorporate measurement error in dynamic modeling is to add a noise term to the model, typically white noise, to represent the measurement error. White noise is a series of random variables that are identically and independently distributed (Chatfield, 2004). For the AR model with measurement error (AR(1)+WN), the white noise ω_t is simply added to each observation y_t (see **Figure 1B**). We assume that this white noise is normally distributed with a mean of zero and variance σ_ω^2. This results in the following model specification for the AR(1)+WN model

$$\begin{aligned} y_t &= \mu + \tilde{y}_t + \omega_t \\ \tilde{y}_t &= \phi \tilde{y}_{t-1} + \epsilon_t \quad (3) \\ \epsilon_t &\sim N\left(0, \sigma_\epsilon^2\right) \quad (4) \\ \omega_t &\sim N\left(0, \sigma_\omega^2\right). \quad (5) \end{aligned}$$

Important to note is that when ϕ is equal to zero, the measurement error and dynamic error will no longer be discernible from each other, because they are only discernible from each other from the merit that the innovations are passed to future time points through ϕ, while the measurement errors are not. In that case, the AR(1)+WN model is no longer identified, which is problematic for estimating the model parameters. Further note that when ϕ is nonzero, the higher $|\phi|$, the easier it will be to discern measurement error from the innovations, and as such the model will be easier to identify empirically, and likely easier to estimate. Hence, in this sense the (empirical) identification of the AR(1)+WN model may be seen as dimensional rather than dichotomous, ranging from unidentified when ϕ is zero, to maximally empirically identified when $|\phi|$ is one.

2.4. Incorporating Measurement Error: The ARMA(1,1) Model

Another way to incorporate measurement error into an AR(1) model that is relatively frequently suggested in the literature on dynamic modeling with measurement error, is to use an autoregressive moving average (ARMA) model (see for instance: Granger and Morris, 1976; Deistler, 1986; Chanda, 1996; Swamy et al., 2003; Wagenmakers et al., 2004; Staudenmayer and Buonaccorsi, 2005; Costa and Alpuim, 2010; Patriota et al., 2010). Granger and Morris (1976) have shown that the AR(p)+WN model is equivalent to an ARMA(p,p) model, where p stands for the number of lags included in the model. As a result, an ARMA(1,1) model can be used as an indirect way to fit an AR(1) model and take measurement error into account (Granger and Morris, 1976; Staudenmayer and Buonaccorsi, 2005; Wagenmakers et al., 2004). One advantage of fitting an ARMA(1,1) model rather than fitting an AR(1)+WN model directly, is that it can be estimated with a wide range of estimation procedures, and a wide range of software, including for instance SPSS. A second important advantage is that the ARMA(1,1) is identified when the value of ϕ is equal to zero, so that in practice it may be easier to estimate than the AR(1)+WN model.

An ARMA(1,1) process consists of an AR(1) process, and a moving average process of order 1 [MA(1)]. In an MA(1) process, the current state $\tilde{y}_t$ depends not only on the innovation, ϵ_t^*, but also on the previous innovation ϵ_{t-1}^*, through moving average parameters θ.[2] For example, consider the daily introverted behavior for a specific person. On a certain day, the person has a shameful experience, resulting in a strong boost (e.g., an innovation or perturbation) in introverted behavior. The next day, this person may display lingering heightened introverted behavior from the previous day as a result of an AR effect, but there may also be a delayed response to the perturbation from yesterday, for instance because the person remembers the events of the previous day. The strength of the delayed response depends on the size of θ. The ARMA(1,1) model, which is depicted in **Figure 1C**, can be specified as:

$$\begin{aligned} y_t &= \mu + \tilde{y}_t \\ \tilde{y}_t &= \phi \tilde{y}_{t-1} + \theta \epsilon_{t-1}^* + \epsilon_t^* \quad (6) \\ \epsilon_t^* &\sim N\left(0, \sigma_\epsilon^{2*}\right). \quad (7) \end{aligned}$$

The ARMA(1,1) model is characterized by four parameters, that is, the mean μ, AR parameter ϕ, moving average parameter θ, and innovation variance σ_ϵ^{2*}. The model is stationary when ϕ lies between –1 and 1, and is invertible if θ lies between -1 and 1 (Chatfield, 2004; Hamilton, 1994).

If the true underlying model is an AR(1)+WN model, the ϕ and μ parameter in an ARMA(1,1) will be equal to those of the AR(1)+WN model. Granger and Morris (1976) have shown that the innovation variance σ_ϵ^2 and measurement error variance σ_ω^2 can be calculated from the estimated θ, ϕ, and σ_ϵ^{2*} as follows (see also Staudenmayer and Buonaccorsi, 2005),

$$\begin{aligned} \sigma_\omega^2 &= (-\phi)^{-1}\theta\sigma_\epsilon^{2*}, \quad (8) \\ \sigma_\epsilon^2 &= (1+\theta^2)\sigma_\epsilon^{2*} - (1+\phi^2)\sigma_\omega^2. \quad (9) \end{aligned}$$

It is important to note that while the AR(1)+WN models is equivalent to an ARMA(1,1) model, an ARMA(1,1) models is not necessarily equivalent to an AR(1)+WN model. That is, it is only possible to transform the ARMA(1,1) parameters to AR(1)+WN model parameters under these restrictions in line with an underlying AR(1)+WN model (Granger and Morris, 1976; Staudenmayer and Buonaccorsi, 2005):

[2]We add the * to ϵ, to distinguish the innovations for the ARMA(1,1,) model from the innovations of the AR(1)+WN model.

$$\frac{1}{1+\phi^2} > \frac{\theta}{1+\theta^2}(-\phi^{-1}) \geq 0 \tag{10}$$

3. Simulation Study Methods

We present a simulation study in which we simulate data according to an AR process with added measurement error. We fit an AR(1) model to the data in order to illustrate the effects of ignoring any present measurement error, and compare the performance of the AR(1) model to the AR(1)+WN, and ARMA(1,1) model, which both account for measurement error. Furthermore, we will compare the performance of the Bayesian and frequentist estimation of these models.

3.1. Frequentist Estimation

For the frequentist estimation of the AR(1) model and the ARMA(1,1) model a relatively wide range of procedures and software is available. Potential estimation procedures for fitting the AR(1)+WN model include specially modified Yule-Walker equations, and modified Least Squares estimation procedures (Chanda, 1996; Staudenmayer and Buonaccorsi, 2005; Dedecker et al., 2011). However, we opt to use the (linear, Gaussian) state-space model, for which the Kalman Filter (Harvey, 1989; Kim and Nelson, 1999) is used to estimate the latent states, while Maximum Likelihood is used to estimate the model parameters (c.f., Staudenmayer and Buonaccorsi, 2005, for this approach, but with the measurement error variance considered as known). This is an especially convenient modeling framework for psychological longitudinal modeling, as it readily deals with missing data, and is easily extended to multivariate settings, or to include latent variables (c.f., Hamilton, 1994; Harvey, 1989; Kim and Nelson, 1999).

In the state-space model representation, a vector of observed variables is linked to a vector of latent variables—also referred to as "state variables"—in the *measurement equation*, and the dynamic process of the latent variables is described through a first-order difference equation in the *state equation* (Hamilton, 1994; Harvey, 1989; Kim and Nelson, 1999). That is, the measurement equation is

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{d} + \boldsymbol{F}\tilde{\boldsymbol{y}}_t + \boldsymbol{\omega}_t \\ \boldsymbol{\omega}_t &\sim MvN\left(\boldsymbol{0}, \boldsymbol{\Sigma}_\omega\right), \end{aligned} \tag{11}$$

where $\boldsymbol{y}_t$ is an $m \times 1$ vector of observed outcome variables, $\tilde{\boldsymbol{y}}_t$ is an $r \times 1$ vector of latent variables, $\boldsymbol{d}$ is an $m \times 1$ vector with intercepts for the observed variables, $\boldsymbol{F}$ is an $m \times r$ matrix of factor loadings, and $\boldsymbol{\omega}_t$ is an $m \times 1$ vector of residuals that are assumed to be multivariate normally distributed with zero means and $m \times m$ covariance matrix $\boldsymbol{\Sigma}_\omega$. The state equation (also referred to as the transition equation) is specified as

$$\begin{aligned} \tilde{\boldsymbol{y}}_t &= \boldsymbol{c} + \boldsymbol{A}\tilde{\boldsymbol{y}}_{t-1} + \boldsymbol{\epsilon}_t \\ \boldsymbol{\epsilon}_t &\sim MvN\left(\boldsymbol{0}, \boldsymbol{\Sigma}_\epsilon\right), \end{aligned} \tag{12}$$

where $\boldsymbol{c}$ is an $r \times 1$ vector of intercepts for the latent variables, $\boldsymbol{A}$ is an $r \times r$ matrix of structural coefficients, and $\boldsymbol{\epsilon}_t$ is an $r \times 1$ vector of residuals, which are assumed to be multivariate normally distributed with zero means and $r \times r$ covariance matrix $\boldsymbol{\Sigma}_\epsilon$.

The previously discussed AR(1) and AR(1)+WN model are both already specified in terms of a state-space representation in Equations (1) through (5) (simplified where possible). For the state-space model specification for the ARMA(1,1) model vector $\boldsymbol{d}$ is μ, $\boldsymbol{F}$ is $[\,1\ 0\,]^T$, $\tilde{\boldsymbol{y}}_t$ is $[\,\tilde{y1}_t\ \tilde{y2}_t\,]^T$, $\boldsymbol{\Sigma}_\omega$ is a zero matrix, $\boldsymbol{c}$ is a zero vector, $\boldsymbol{A}$ is 2×2 matrix $\left[\begin{smallmatrix}\phi & 0\\ 1 & 0\end{smallmatrix}\right]$, and 2×2 matrix $\boldsymbol{\Sigma}_\epsilon$ is equal to $\boldsymbol{H}^T\boldsymbol{H}$ with $\boldsymbol{H}$ equal to $[\,\sigma_{1\epsilon^*}\ \theta\sigma_{1\epsilon^*}\,]$, where superscript T indicates the transpose.

To fit the frequentist state-space models we use R, with R packages FKF (Kalman Filter; Luethi et al., 2010) combined with R base package optim (for maximum likelihood optimization; R Development Core Team, 2012). Within optim we used optimization method l-bfgs-b, with lower bounds and upper bounds for ϕ and θ of -1 and 1, -Inf and Inf for μ, and 0 and Inf for σ^2_ϵ, σ^2_ω, and σ^2_ν.

3.2. Bayesian Estimation

Bayesian modeling shares a lot of conveniences with the frequentist state-space modeling framework: For instance, like frequentist state-space modeling procedures, Bayesian modeling can deal conveniently with missing data, is flexible in modeling multivariate processes, and in including latent variables in the model. Particular to Bayesian modeling is the relative ease in extending models to a hierarchical or multilevel setting (e.g., Lodewyckx et al., 2011; De Haan-Rietdijk et al., 2014). Another advantage may be the possibility to include prior information in the analysis, based, for instance, on expert knowledge or results from previous research (e.g., Rietbergen et al., 2011, 2014). Finally, the Bayesian estimation procedures are not dependent on large sample asymptotics like the frequentist procedures, and may therefore perform better for smaller samples (Dunson, 2001; Lee and Wagenmakers, 2005). Because currently there is no literature on the Bayesian estimation performance for the AR(1)+WN model, we will compare the performance of the Bayesian AR(1), ARMA(1,1), and AR(1)+WN model with the frequentist modeling equivalents in a simulation study.

In Bayesian estimation the information in the data, provided through the likelihood, is combined with a prior distribution using Bayes' rule (c.f., Gelman et al., 2003; Hoijtink et al., 2008). The prior distribution is specified such that it contains prior information the researcher would like to include in the analysis. Here we prefer to specify uninformative prior distributions that contain minimal prior information, such that their influence is minimal. Specifically, we use the following prior specifications across the three models: A *uniform*(0, 20) prior on σ^2_ω, σ^2_ϵ, and σ^2_ν, a *uniform*$(-1, 1)$ prior on ϕ and θ, and a *normal*$(0, 0.001)$ prior for μ (specified with precision rather than variance). When the prior distribution and the likelihood are combined using Bayes' rule, this results in the posterior probability distribution or density of the estimated parameters. Summary statistics based on this distribution can then be used to summarize the information on the estimated parameters, for instance, the mean or median may be used to obtain a point estimate for an estimated parameter, and the posterior standard deviation can be used to describe the uncertainty around that point estimate.

Although it is possible to obtain the posterior distribution analytically for some simple models, the Bayesian estimation

of more complex models is usually done with Markov Chain Monte Carlo algorithms, such as Gibb's sampling, which relies on consecutively samples from the conditional distributions of the parameters (rather than directly from their joint distribution, c.f., Casella and George, 1992); when the procedure has converged, one effectively samples from the (joint) posterior distribution. These samples can then be used as an approximation of the underlying posterior distribution, which in turn can be used to obtain point estimates for the parameters. A particularly desirable feature of MCMC procedures is that, based on the samples of the estimated parameters, it is also possible to calculate new statistics and obtain their posterior distribution. For instance, based on the estimated parameters θ, ϕ, and σ_ϵ^{2*} for the ARMA(1,1) model, we will calculate the innovation variance σ_ϵ^2 and measurement error variance σ_ω^2 in each sample, such that we obtain posterior distributions for these parameters. In our simulations we use the free open source software JAGS (Plummer, 2003) which employs a Gibb's sampling algorithm, in combination with the R package Rjags (Plummer et al., 2014).

3.3. Simulation Conditions

Throughout the simulation study, we simulated 1000 data sets per condition according to the AR(1)+WN model specified in Equations (3–5) using R (R Development Core Team, 2012). For all conditions, the mean of the model is fixed to 2. The study consists of three parts. First, we examine the effect of *the proportion of measurement error variance to the total variance*, on parameter recovery. The total variance for the AR(1)+WN is the sum of the variance for an AR(1) model and the measurement error variance: $\sigma_{total}^2 = \sigma_\epsilon^2/(1-\phi^2) + \sigma_\omega^2$ (c.f., Harvey, 1989; Kim and Nelson, 1999). To vary the proportion of σ_ω^2 to the total variance, ϕ and σ_ϵ^2 are both fixed to 0.5 in this study while the measurement error variance is varied. Specifically, the measurement error variance takes on the values 0, 0.1, 0.2, 0.3, 0.5, 0.7, 1, 2, 4, and 12, which results approximately in the following proportions of measurement error variance to the total variance: 0, 0.13, 0.23, 0.31, 0.43, 0.51, 0.6, 0.75, 0.86, and 0.95.

Second, we examine the effect of *the size of* ϕ on parameter recovery. We vary ϕ over the values −0.75, −0.5, −0.25, 0, 0.25, 0.5, and 0.75. The proportion of measurement error variance to the total variance of the AR(1)+WN process is fixed to 0.3 here, through varying the innovation variances σ_ϵ^2 by approximately 1.2, 1.1, 0.9, 0.5, 0.9, 1.1, and 1.2 respectively.

Third, we examine the effects of *sample size*. In part 1 and 2 of the study we use a sample size 100 repeated measures. We based this number roughly on what one may expect for research in psychology: Typically, what we see in time series applications in psychology is a range of about 60–120 repeated measures per person (e.g., see Nezlek and Gable, 2001; Rovine and Walls, 2006; Madhyastha et al., 2011; Ferrer et al., 2012; Wang et al., 2012; Adolf et al., 2015). However, in preliminary analyses we found difficulties in estimating the model with a small sample size, especially for the frequentist estimation procedure, that pointed to empirical underidentification (we elaborate on this in the next section). Therefore, we varied sample size by 100, 200, and 500. For this part of the study σ_ϵ^2, σ_ω^2, and ϕ were fixed to 0.5, implying a proportion of measurement error variance to the total variance of 0.43.

We judge the performance of each model based on: (a) its bias in the estimates; (b) the absolute error in the estimates; and (c) coverage rates for the 95% confidence or credible intervals. It is not clear whether Bayesian 95% credible intervals should have exactly 95% coverage rates, however, with uninformative priors we would expect this to be the case. Moreover, we consider it informative to see how often the true value lies within the credible interval across multiple samples (e.g., if this occurs very rarely this seems problematic for making inferences).

For the coverage rates of the variances estimated with the frequentist ML procedure, we calculate the confidence intervals based on a χ^2 distribution with $n-1$ degrees of freedom as follows: $CI(\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{\alpha/2}})$, where n is the sample size, and s^2 is the estimated variance.

3.4. Expectations

For part 1, we expect that all models will decrease in performance (i.e., more bias and absolute error, lower coverage rates) as the proportion of measurement error variance increases, because an increase in random noise should make it harder to distinguish an (autoregressive) effect. Furthermore, we expect that the decrease in performance will be larger for the AR(1) model than for the ARMA(1,1) and AR(1)+WN model. Specifically, based on Staudenmayer and Buonaccorsi (2005), we expect a bias in the estimates of ϕ in the AR(1) model of approximately 0, −0.07, −0.12, −0.16, −0.21, −0.26, −0.30, −0.38, −0.43, and −0.47, given that the proportions of measurement error variance are 0, 0.13, 0.23, 0.31, 0.43, 0.51, 0.6, 0.75, 0.86, and 0.95.

For part 2, we expect that the AR(1)+WN and ARMA(1,1) models will improve in performance as the value of $|\phi|$ increases, given that σ_ω^2 and σ_ϵ^2 should be more easily distinguished from each other as $|\phi|$ approaches 1. We are specifically interested in the performance of the AR(1)+WN model compared to the ARMA(1,1) model when $|\phi|$ is relatively small. Given that the ARMA(1,1) model is identified regardless of the value of ϕ, we expect the ARMA(1,1) model may converge better, and therefore to perform better when ϕ is relatively close to zero than the AR(1)+WN model, which is no longer identified when ϕ is equal to zero.

For part three, we expect that performance will improve as sample size increases for the ARMA(1,1) model and the AR(1)+WN model, both in the frequentist and Bayesian estimation procedure. Finally, we expect that the Bayesian procedure will perform better than the frequentist state-space procedures for smaller sample sizes, given that both modeling procedures have similar benefits, but the Bayesian estimation procedure is not dependent on large sample asymptotics (Dunson, 2001; Lee and Wagenmakers, 2005).

4. Simulation Study Results

In this section we present the results of the simulation study. As was mentioned before, for a sample size of 100 we found some convergence issues especially for the frequentist ML procedure.

Given that convergence is an important precondition for obtaining reasonable parameter estimates, we start by discussing the convergence of the Bayesian models and frequentist models across the different parts of the simulation study. After that, we discuss the parameter recovery performance for each condition specific for each of the three parts of the simulation study. We end with a summarizing conclusion.

4.1. Convergence of the Bayesian Procedures

For the Bayesian procedures we obtained three chains of 40,000 samples each for each replication, half of which was discarded as burn-in. We judged convergence based on the multivariate Gelman-Rubin statistic and autocorrelations for all replications, and we inspected the mixing of the three chains visually a number of replications (c.f., Gelman and Rubin, 1992; Brooks and Gelman, 1998). For the AR(1) model the chains mixed well, the Gelman Rubin statistic was generally equal to one, and the autocorrelations for the parameters decreased exponentially across all conditions.

For the ARMA(1,1) the chains generally mixed well, and the Gelman Rubin statistic was equal to one across all conditions.[3] The autocorrelations for the parameters decreased slower than for the AR(1) model, and decreased most slowly when the proportion of measurement error variance was higher than 50% or $|\phi|$ was zero.

For the AR(1)+WN model, overall the chains mixed well and the Gelman Rubin Statistic was equal to one for most replications. For approximately 1–2% of the data sets the Gelman Rubin statistic was larger than 1.1, indicating possible non-convergence, with the exception of the condition where $\phi = 0.75$, for which it was 8%. Closer inspection indicated that these problems usually originated and were limited to μ. The percentage of non-convergence is larger for the condition $\phi = 0.75$, most likely because when ϕ is strong and positive it is most difficult to estimate μ because observations may tend to linger longer above or below the mean. The autocorrelations for the AR(1)+WN model are higher overall, and slower to decrease than those for the AR(1) and ARMA(1,1) model across all conditions. More measurement error and a closer $|\phi|$ to zero, was associated with more slowly decreasing autocorrelations.

4.2. Convergence of the (Frequentist) ML with State-space Modeling Procedures

For the ML procedure we encountered three types of problems: (1) negative standard errors for the estimated parameters , (2) optim failing to initialize (more rarely), and (3) Heywood cases (negative variances) for the measurement error variance or the innovation variance. The first and second type of problem could usually be resolved by providing alternative starting values and rerunning the model. For a small percentage of data sets, five sets of starting values still did not resolve these issues (for the number of data sets per condition, see Table A1 in Supplementary Materials). These data sets are excluded from the parameter recovery results. When sample size was increased to 200 or 500 repeated measurements, these problems were no longer encountered.

The third type of problem—Heywood cases—was much more prevalent, and could generally not be resolved by providing different starting values. For the AR(1)+WN model, for 10–55% of the replications σ^2_ω, or more rarely σ^2_ϵ, were estimated at the lower bound of zero. For the ARMA(1,1) model, we similarly detected Heywood cases for σ^2_ω and σ^2_ϵ (note that σ^2_ω and σ^2_ϵ are calculated a posteriori based on the estimated ϕ, θ and σ^{2*}_ϵ by means of Equations 8 and 9). In the case that for the AR(1)+WN model σ^2_ω or σ^2_ϵ were estimated at the lower bound, usually a Heywood case would also observed for the ARMA(1,1) model for that replication. The proportions of Heywood cases for σ^2_ω and σ^2_ϵ across all conditions are reported in Table A1 in the Supplementary Materials.

The number of Heywood cases increased when: (1) ϕ got closer to zero, such that it is harder to discern measurement errors from innovations (2) when there was very little measurement error, such that σ^2_ω was already close to zero, and (3) There was a lot of measurement error, such that all parameter estimates were uncertain (large standard errors). This indicates issues of empirical identification, and as such we expected these issues to decrease as sample size increases.

The Heywood cases for σ^2_ϵ and σ^2_ω decreased as sample size increased—however, the issues were not resolved completely: For $n = 200$ almost 30% of the data sets still returned a Heywood case, and for $n = 500$ almost 13% still returned a Heywood case. Given that for smaller sample sizes (e.g., less than 500), which are much more common in psychological studies, the proportion of replications with Heywood cases was quite large for many conditions, this seems quite problematic. In practice, encountering such a result may lead a researcher to erroneously conclude that there most likely is no considerable measurement error variance, so that a regular AR(1) model should suffice.

In the following sections, where we discuss the parameter recovery results, the data sets with Heywood cases for σ^2_ω or σ^2_ϵ are included in the results, because to exclude so many data sets would make a fair comparison to the Bayesian procedure (for which no data sets need to be excluded) problematic. However, the results with these data sets excluded for the ML AR(1)+WN model and ARMA(1,1) model are presented and discussed in Supplementary Materials. Finally note that, in contrast to our expectations, in the ML procedure the ARMA(1,1) model does not seem to converge more easily than the AR(1)+WN model. In general it seems that in order to properly estimate and distinguish the measurement error variance from the innovation variance using ML, quite large sample sizes are required.

4.3. Parameter Recovery for Different Proportions of Measurement Error

In general, as the proportion of measurement error increases, the estimated parameters become increasingly more biased, the absolute errors become larger, and coverage rates become

[3] By visually inspecting the chains for μ in the ARMA(1,1) model, we found some extreme values for some of the Gibb's samples (visible as large "spikes" in the chains). To limit these extreme values we adjusted the normal prior for μ to have a smaller variance (10), however this did not resolve the issue completely. As a result, the posterior standard deviation for μ was very large, however, the effects on the point estimates and credible intervals seem limited when we compare these results for μ to those of the other models.

lower, as expected. In **Figure 2** we provide plots of the 95% coverage, absolute errors, and bias for each model, condition, and parameter. As can be seen from this figure, overall, the Bayesian AR(1)+WN model outperforms the other procedures in terms of coverage rates and absolute errors, and for the variance parameters also in terms of bias. The ML state-space AR(1)+WN model performs second-best overall, and performs the best for ϕ in terms of bias. The Bayesian and frequentist AR(1) and ARMA(1,1) models perform relatively poorly in all respects. However, the ARMA(1,1) models result in better coverage rates for ϕ than the AR(1) models, so that an ARMA(1,1) model is still preferred over a simple AR(1) model. Below, we will discuss the results in more detail, per parameter.

For μ, all models perform similarly well in terms of bias and absolute error, as can be seen from the top-left panel of **Figure 2**. In terms of coverage rates, the Bayesian AR(1) and AR(1)+WN model outperform the other models for μ, most pronouncedly when the proportion of measurement error is high.

For ϕ, the models that perform the best in terms of bias are the ML AR(1)+WN model, followed by the Bayesian AR(1)+WN model (see the top-right panel in **Figure 2**). The bias for ϕ in both AR(1) models is in line with our expectations, increasing from approximately 0 to −0.5 as measurement error increases. As can be seen from the top-right panel of **Figure 2**, in terms of absolute error for ϕ, the Bayesian AR(1)+WN model performs the best, followed by the ML AR(1)+WN model. The top-right panel of **Figure 2** shows that the coverage rates for ϕ

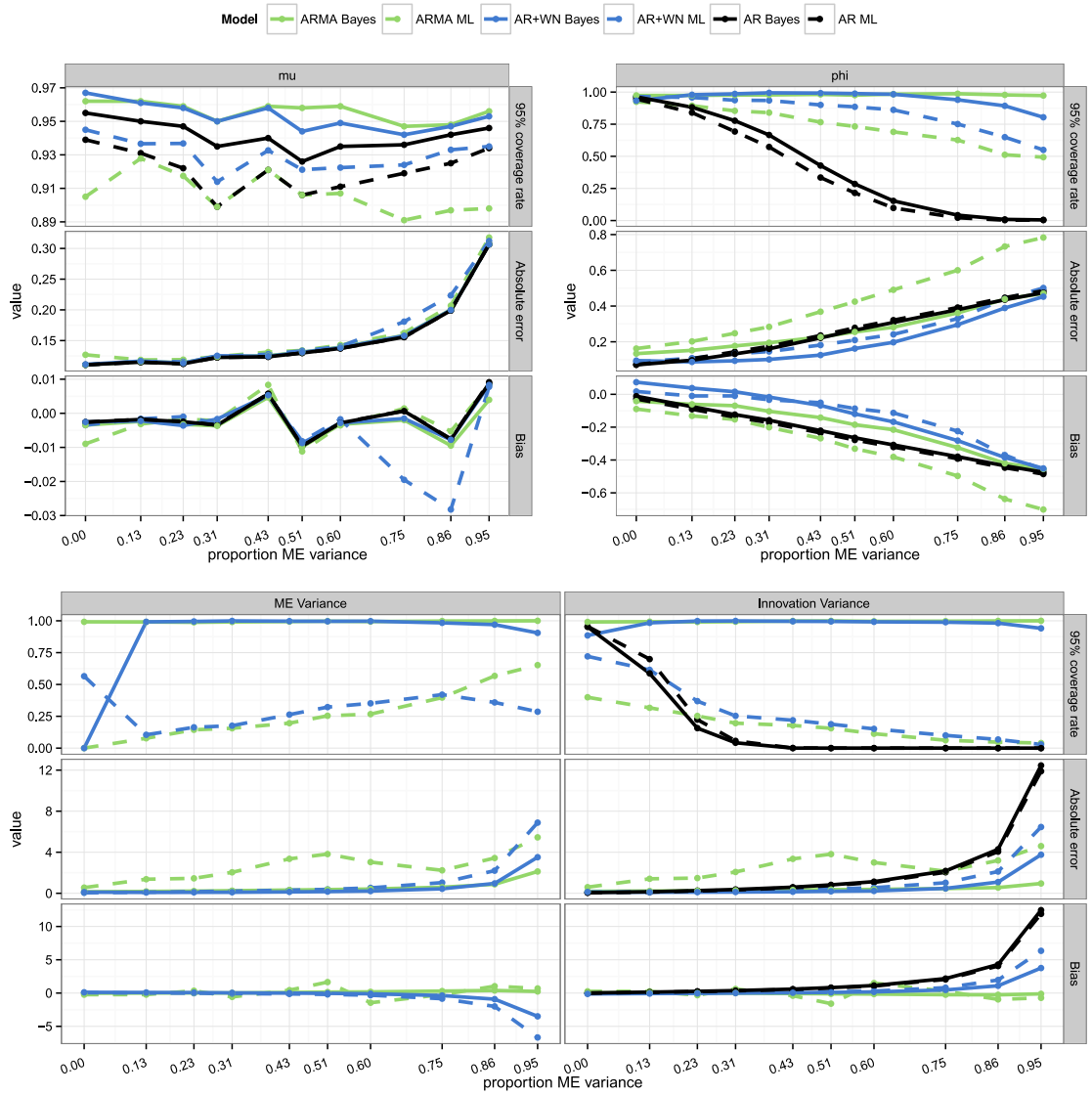


FIGURE 2 | Coverage rates, absolute errors, and bias for the parameter estimates for the frequentist and Bayesian AR(1), ARMA(1,1), and AR(1)+WN models across different proportions of measurement error variance to the total variance.

based on the 95% CI's for the Bayesian ARMA(1,1) model are consistently higher than those for the other models, however, this is a result of having wider credible intervals, rather than a result of more precise estimates for ϕ. The coverage rates for the Bayesian AR(1)+WN model are most stable across the different proportions of measurement error variance. The coverage rates for this Bayesian model are generally higher than 0.95[4], only dropping below 0.95 when 75% or more of the total variance is measurement error variance. In comparison, the ML AR(1)+WN model starts with a coverage rate of approximately 0.95 for ϕ when measurement error is absent, and the coverage decreases as measurement error increases (with a lowest coverage of 0.55 when 95% of the variance is due to measurement error). The ML ARMA(1,1) model and the Bayesian and ML AR(1) models perform the worst, as can be seen from **Figure 2**. Note that for the AR(1) models, the coverage rates for ϕ are already below 90% when the proportion of measurement error variance is as little as 0.13.

In the bottom panel of **Figure 2** the results for σ_ω^2 and σ_ϵ^2 are displayed. When the proportion of error variance is larger than about 0.3, the Bayesian AR(1)+WN model starts to outperform the ML AR(1)+WN model in terms of bias for σ_ω^2 and σ_ϵ^2. Further, it can be seen from **Figure 2** that for the AR(1)+WN models, when the proportion of measurement error is small, the measurement error variance is slightly overestimated, while when the proportion of measurement error is large, the measurement error variance is underestimated. The coverage rates are the highest for the Bayesian AR(1)+WN and ARMA(1,1) model. Note that for the ARMA(1,1) model σ_ω^2 and σ_ϵ^2 are calculated based on the estimated ARMA(1,1) parameters. For the Bayesian model this was done in each Gibbs sample by means of Equations (8) and (9), resulting in a posterior distribution for σ_ω^2 and σ_ϵ^2. However, depending on the specific values of the ARMA(1,1) parameters in each Gibbs sample, σ_ω^2 and σ_ϵ^2 may become quite large or even negative. As a result, the posterior standard deviations and credible intervals for σ_ω^2 and σ_ϵ^2 in the Bayesian ARMA(1,1) model can be quite large, including negative and large positive values. The confidence intervals for the variances parameters in frequentist procedures are consistently too narrow, which results in low coverage rates, as can be seen from the bottom panel of **Figure 2**. As such, for the two variances, the Bayesian AR(1)+WN model performs best in terms of coverage rates, followed by the Bayesian ARMA(1,1) model (which has higher coverage rates, but much wider intervals), and the ML AR(1)+WN model. The same pattern holds for the absolute errors as can be seen in **Figure 2**.

4.4. Parameter Recovery for Different Values of ϕ

For this part of the study, the value of ϕ was varied from -0.75 to -0.5, -0.25, 0, 0.25, 0.5, and 0.75. As can be seen from the top-left panel of **Figure 3**, for μ all the models perform very similarly in terms of bias, absolute errors, and coverage rates. The absolute errors and bias increase as ϕ becomes larger, because when ϕ is strong and positive, observations may tend to linger longer above or below the mean than when ϕ is weak or negative, making it harder to estimate μ.

As can be seen from the top-right and bottom panel of **Figure 3**, the results for ϕ and the variance parameters are symmetric for negative and positive values of ϕ (or mirrored in the case of bias). As such, we will discuss these results in terms of $|\phi|$. For the parameters ϕ, σ_ϵ^2 and σ_ω^2, performance increases as $|\phi|$ increases, except the AR(1) models, for which it is the opposite. Overall, the Bayesian AR(1)+WN performs best, followed by respectively the ML AR(1)+WN model, the Bayesian ARMA(1,1) model, and the ML ARMA(1,1) model. The performance of the latter three models decreases considerably more as $|\phi|$ decreases than that of the Bayesian AR(1)+WN model, as can be seen from **Figure 3**.[5] For the two variances, the ML AR(1)+WN model outperforms the Bayesian model in terms of bias. Finally, we find that when $|\phi|$ is relatively close to one, the measurement error variance is underestimated, however, when $|\phi|$ is relatively small, the measurement error variance was actually overestimated, as can be seen from the bottom panel of **Figure 3**.

4.5. Parameter Recovery for Different Sample sizes

For this part of the simulation study, the sample size was varied from 100 to 200 and 500. As shown in **Figure 4**, as sample size increases, parameter recovery improves: Bias and absolute errors decrease, while coverage rates become closer to 0.95. We Further, the ML AR(1)+WN results become more similar to those of the Bayesian AR(1)+WN model as sample size increases, although the Bayesian model still outperforms the ML model in terms of absolute error and coverage: The Bayesian procedure results in higher coverage rates, but less wide intervals, that is, in more precise estimates than the ML procedure for ϕ. Note that the performance of the ML and Bayesian ARMA(1,1) models only near the performance of the AR(1)+WN models as sample size has increased to 500 observations.

4.6. Conclusion

Overall, the Bayesian AR(1)+WN model performs better than the other five procedures we considered. We expected that the ARMA(1,1) models may outperform the AR(1)+WN models in parameter recovery, because we expected this model to have less trouble with identification and convergence. Interestingly, although the Bayesian ARMA(1,1) model seems to converge more easily than the Bayesian AR(1)+WN model, the AR(1)+WN model still outperforms the ARMA(1,1) model in terms of parameter recovery, even when ϕ is close or equal to zero. The ML AR(1)+WN model and ARMA(1,1) models are both unstable for small sample sizes ($n = 100$), frequently resulting in Heywood cases for the innovation and measurement error variances. However, the ML AR(1)+WN model still

[4]While it may seem undesirable that the Bayesian model has "too high" coverage rates, indicating too large credible intervals or exaggerated uncertainty about the estimated parameters, it is important to note that compared to the ML model, the Bayesian estimates actually have smaller posterior standard deviations than the ML standard errors for ϕ.

[5]The diverging patterns in the bias and absolute errors for the ML ARMA(1,1) model is a result of the Heywood cases discussed in Section 4.2; when the Heywood cases are removed the pattern is similar to the patterns of the other models, as can be seen in Figures B1,B2 in Supplementary Materials.

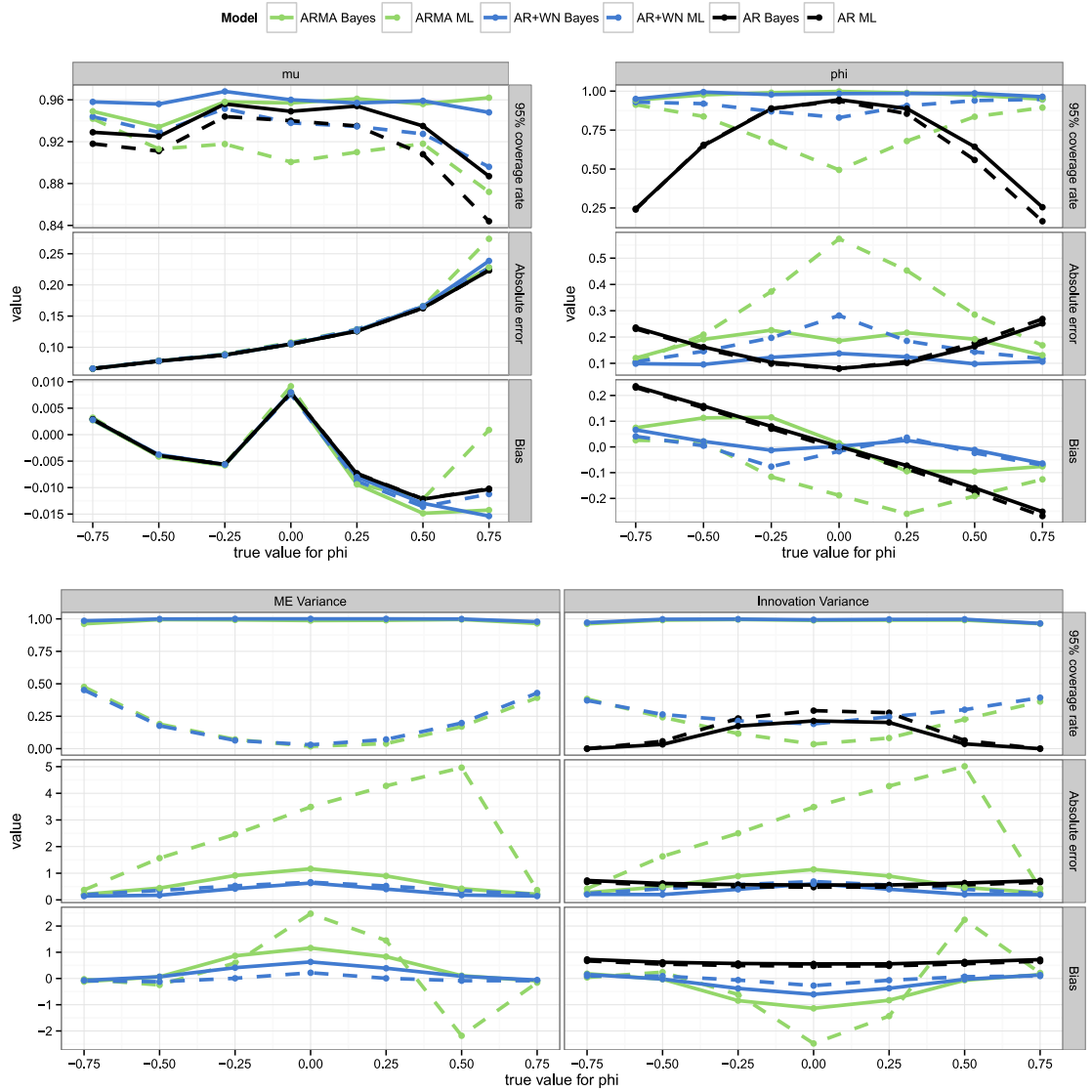


FIGURE 3 | Coverage rates, absolute errors, and bias for the parameter estimates for the frequentist and Bayesian AR(1), ARMA(1,1), and AR(1)+WN models across different values for ϕ.

performs relatively well for estimating ϕ compared to the AR(1) models. For a smaller sample size of 100 observations the Bayesian procedure outperforms the frequentist ML procedure. When sample sizes are larger, the discrepancies between the Bayesian and frequentist AR(1)+WN model decrease, although the confidence intervals for the variance parameters in the frequentist procedures are consistently too narrow. As expected, the AR(1) models severely underestimate $|\phi|$, which is reflected in large bias and absolute errors, and low coverage rates. Finally, we note that although the AR(1)+WN models perform considerably better than the AR(1) models, some bias in ϕ still remains, because the innovations and measurement errors cannot be perfectly discerned from each other. Generally, the more measurement error and the lower $|\phi|$, the more the estimate of $|\phi|$ will be biased, even when measurement error is taken into account by the model.

5. Empirical Application on Mood Data

To further illustrate the AR(1), ARMA(1,1), and AR(1)+WN model discussed above, we make use of time series data that was collected from female first year social science students at Utrecht University in 2007. Eleven women kept a daily electronic diary for approximately 3 months (across participants the minimum was 90 observations, the maximum 107 observations), in which they filled out how they felt that day on a scale from 1 to 100—1 meaning worst ever, and 100 meaning best ever. Three of the eleven women were excluded from the current study

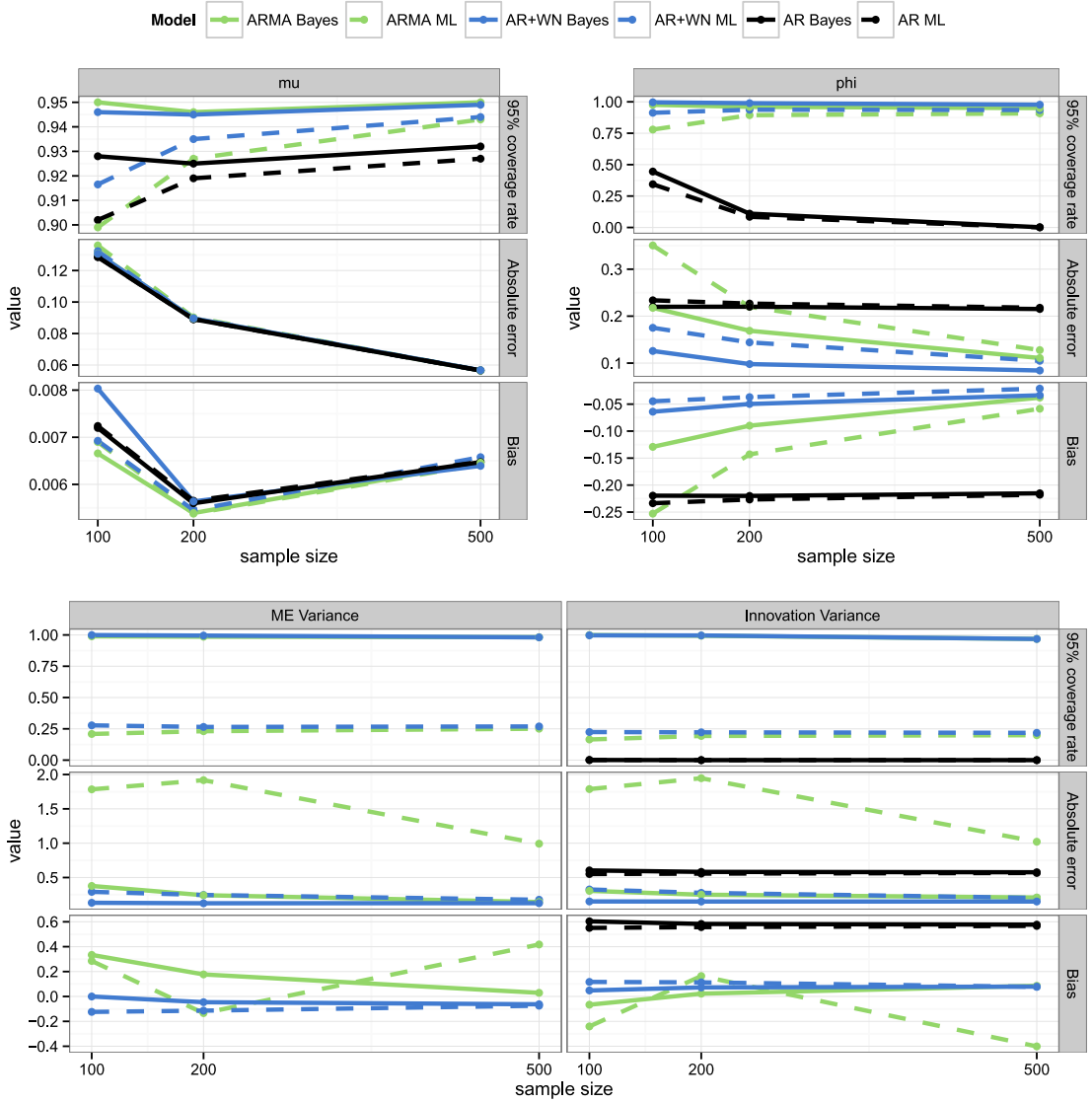


FIGURE 4 | Coverage rates, absolute errors, and bias for the parameter estimates for the frequentist and Bayesian AR(1), ARMA(1,1), and AR(1)+WN models across sample sizes.

because of non-compliance, issues with the electronic devices, and one woman had very little variation in her scores. For the remaining women the average number of missing observations was approximately nine. Values for these missing observations will be automatically imputed as part of the estimation procedure, based on the specified model.

We are interested in finding out to what extent current mood influences mood the following day. As such, we are interested in fitting an AR(1) model, and specifically in the AR effect reflected in parameter ϕ. However, the mood of each person is not likely to be perfectly measured. For instance, it is possible that participants accidentally tapped the wrong score when using the electronic diary stylus to fill in the questionnaire. Furthermore, the participants evaluate their mood for the day on average, such that momentary influences around the time of filling out the diary may have colored their evaluation of the whole day (i.e., a form of retrospective bias). In fact, anything that is not explicitly measured and modeled, and of which the influence does not carry-over to the next day, can be considered measurement error. As such, it seems likely that there is at least some measurement error present in the data. Therefore, we fit the AR(1)+WN model to take this measurement error into account, and for illustrative purposes compare it to an ARMA(1,1) model, and an AR(1) model (which disregards measurement error). The data and codes for running the analyses are included in the Supplementary Materials. We make use of a Bayesian modeling procedure, given that the results from our simulation study indicate that the parameter recovery performance of the Bayesian procedure is

better and more stable for this number of repeated measures. The priors we use for the models are aimed to be uninformative, specifically: A *uniform*(0, 500) prior distribution for all variance parameters, a *uniform*(−1, 1) prior distribution for ϕ and θ, and a *normal*(0, 0.001) prior distribution for μ (specified with a precision rather than a variance).

We evaluated the convergence of the AR(1), ARMA(1,1), and AR(1)+WN model by visually inspecting the mixing of the three chains, the Gelman Rubin statistic, and the autocorrelations. For the AR(1) and AR(1)+WN model the chains mixed well, the Gelman Rubin statistic was approximately equal to one, and the autocorrelations for the parameters decreased within 50–100 lags across all participants. For the ARMA(1,1) model this was the case, except for participants 3 and 8.[6] We included the ARMA(1,1) estimates for these participants in **Table 1**, but these should be interpreted with caution.

The parameter estimates of the mean μ, AR parameter ϕ, innovation variance σ_ϵ^2, measurement error variance σ_ω^2, and moving average parameter θ for each person are presented in **Table 1**. For most of the eight individuals, the baseline mood is estimated to be around 60–70, which indicates that on average they are in moderately good spirits. Further, we see that across models and persons, the AR parameters are either estimated to be positive, or nearly zero. Participant 8 has an AR effect near zero in both the AR(1) model and the AR(1)+WN model, so that for her, everyday seems to be a "new day": How she felt the previous day does not predict her overall mood today. On the other hand, for participants 2, 4, 5, and 6, the credible intervals for ϕ include only positive values across models: how they feel today depends in part on how they felt yesterday. For the remaining individuals,

[6]For participants 3 and 8 we found that the estimates for ϕ and θ in the ARMA(1,1) model were very dispersed, varying across the entire range of −1 to 1, switching from negative to positive values. A density plot of their samples revealed a bimodal distribution for ϕ and θ (with one peak around negative values, and one for positive values): This seems to be some form of label switching, which is indicative of (empirical) under–identification of the ARMA(1,1) model for these two participants.

TABLE 1 | Parameter estimates for the AR(1), ARMA(1,1), and AR+WN model for the mood of eight women, estimated with Bayesian software.

Pp	Model	μ (95% CI)	ϕ (95% CI)	σ_ϵ^2 (95% CI)	σ_ω^2 (95% CI)	$\sigma_\epsilon^{2^*}$ (95% CI)	θ (95% CI)
1	AR1	75 (72, 79)	0.08 (−0.17, 0.32)	166 (122, 235)	–	–	–
	ARMA	76 (72, 81)	0.53 (−0.32, 0.90)	21.34 (− 91, 180)	125 (− 6, 278)	160 (117, 227)	−0.41 (−0.81, 0.29)
	ARWN	76 (72, 79)	0.39 (−0.23, 0.77)	42 (3, 160)	112 (16, 193)	–	–
2	AR1	63 (59, 68)	0.36 (0.13, 0.57)	188 (141, 256)	–	–	–
	ARMA	63 (58, 69)	0.48 (−0.21, 0.97)	103 (−740, 1087)	69 (− 870, 960)	189 (142, 257)	−0.13 (−0.64, 0.49)
	ARWN	63 (58, 68)	0.52 (0.15, 0.84)	101 (20, 208)	77 (7, 184)	–	–
3	AR1	63 (61, 66)	0.21 (0, 0.42)	108 (81, 148)	–	–	–
	ARMA	64 (61, 66)	0.02 (−0.72, 0.81)	−1 (− 288, 251)	109 (− 134, 418)	105 (79, 144)	0.19 (−0.64, 0.95)
	ARWN	64 (61, 67)	0.40 (−0.01, 0.82)	38 (4, 112)	64 (6, 118)	–	–
4	AR1	56 (53, 58)	0.21 (0.01, 0.42)	103 (78, 141)	–	–	–
	ARMA	54 (40, 59)	0.85 (0.35, 0.99)	7 (1, 47)	75 (44, 112)	95 (71, 130)	−0.68 (−0.87, −0.14)
	ARWN	55 (49, 59)	0.69 (0.07, 0.97)	19 (2, 88)	70 (17, 111)	–	–
5	AR1	69 (64, 75)	0.48 (0.28, 0.67)	174 (131, 239)	–	–	–
	ARMA	69 (62, 77)	0.67 (0.20, 0.92)	86 (24, 348)	61 (− 139, 143)	173 (130, 237)	−0.26 (−0.58, 0.24)
	ARWN	69 (62, 77)	0.67 (0.37, 0.91)	90 (27, 190)	66 (6, 140)	–	–
6	AR1	73 (71, 74)	0.27 (0.07, 0.46)	31 (24, 42)	–	–	–
	ARMA	73 (71, 74)	0.18 (−0.43, 0.66)	22 (− 305, 349)	8 (− 314, 339)	31 (24, 42)	0.09 (−0.45, 0.61)
	ARWN	73 (71, 74)	0.33 (0.01, 0.62)	21 (4, 35)	10 (0.51, 30)	–	–
7	AR1	71 (69, 73)	0.08 (−0.13, 0.28)	105 (79, 144)	–	–	–
	ARMA	71 (65, 75)	0.48 (−0.77, 0.99)	7 (− 132, 175)	87 (− 63, 248)	104 (78, 142)	−0.36 (−0.90, 0.77)
	ARWN	71 (68, 74)	0.26 (−0.57, 0.92)	23 (1, 101)	76 (8, 123)	–	–
8	AR1	73 (71, 74)	0.03 (−0.18, 0.24)	59 (44, 80)	–	–	–
	ARMA	73 (71, 74)	−0.22 (−0.81, 0.84)	−5 (− 131, 102)	67 (− 41, 197)	57 (43, 78)	0.31 (−0.98, 0.95)
	ARWN	73 (71, 74)	−0.03 (−0.65, 0.51)	16 (0.35, 61)	42 (2, 70)	–	–

Note that the negative values for in the credible interval for σ_ϵ^2 and σ_ω^2 for the ARMA(1,1) models result, because they are calculated a posterior based on the samples for ϕ, θ, and $\sigma_\epsilon^{2^*}$ based on Equations (8) and (9): It is possible that for certain combinations of these parameters σ_ϵ^2 and σ_ω^2 become negative. For participants 3 and 8 the ARMA(1,1) model did not converge properly, so that these results should be interpreted with caution.

1, 3, and 7, the point estimates for ϕ are also positive, however, the credible intervals including negative and positive values for ϕ.

When we compare the results for the AR(1) model and the AR(1)+WN model, we find that for all participants except participant 8, the AR parameter is estimated to be higher in the AR(1)+WN model: Because the AR(1) model does not take measurement error into account, the AR parameter is estimated to be lower than for the AR(1)+WN model. The extent to which the estimate for ϕ differs across the AR(1) and AR(1)+WN model, differs from person to person. The larger the estimated measurement error variance relative to the total variance, the larger the difference between the estimated ϕ in the AR(1) and AR(1)+WN model. For instance, for participants 4 and 6 their estimates of ϕ in the AR(1) model are quite similar to each other (i.e., 0.21 and 0.27), but because the measurement error variance for participant 4 is estimated to be much larger than that for participant 6 (i.e., 70 vs. 10), her ϕ in the AR(1)+WN model ϕ is also estimated to be larger (i.e., 0.69 vs. 0.33).

Note that the ARMA(1,1) and AR(1)+WN model should not necessarily give the same results: Although the AR(1)+WN model is equivalent to the ARMA(1,1) model, the reverse is not the case. In other words, it is possible that the ARMA(1,1) model captures a different pattern of variation in the data than the AR(1)+WN model, giving different results. However, when we compare the results for the ARMA(1,1) and AR(1)+WN model, we do find fairly similar results for most of the participants (with exception of participants 3 and 8, who had convergence issues for the ARMA(1,1) model), especially for participants 2 and 5. However, a clearly notable difference is that the ARMA(1,1) model has less precise estimates than the AR(1)+WN model, as can be seen from the relatively wide credible intervals for the ϕ parameters in **Table 1**.

Finally, we note that when we calculate the estimated proportion of measurement error variance relative to the total variance based on the AR(1)+WN model for each participant, we find a range of 0.34–0.50 (i.e., 0.36, 0.47, 0.48, 0.50, 0.46, 0.42, 0.46, and 0.34 respectively). This implies that across these eight women, between one third to half of the observed variance is estimated to be due to measurement error.

6. Discussion

In this paper we demonstrate that it is important to take measurement error into account in AR modeling. We illustrated the consequences of disregarding measurement error present in the data both in a simulation study, and an empirical example based on a replicated time series design. Further, we compared the parameter recovery performance for the Bayesian and frequentist AR(1)+WN and ARMA(1,1) models that account for measurement error. Ignoring measurement error present in the data is known to result in biased estimates toward zero of the AR effects in AR(1) models, with the extent of the bias depending on the proportion of measurement error variance and the size of ϕ (Staudenmayer and Buonaccorsi, 2005). Our simulations also demonstrated this bias, and showed large absolute errors and importantly, very poor coverage rates for the AR effect when measurement error is disregarded, regardless of sample size. For research in psychology, for which it is very difficult or perhaps impossible to measure error-free, it seems imperative to consider this potentially large source of variance in our (AR) time series models. In our empirical application for instance, between one third to half of the variance in the data is estimated to be due to measurement error.

Comparing the parameter recovery for the models that incorporate measurement error—the Bayesian and ML ARMA(1,1) model and AR(1)+WN model—revealed that the Bayesian AR(1)+WN model performed best in terms of parameter recovery. It proved relatively tricky to properly estimate the ML ARMA(1,1) and AR(1)+WN model, even for larger sample sizes of 500 repeated measures: These models are prone to Heywood cases in the measurement error variance and to a lesser extent in the innovation variance. This was especially common (up to 55% of the replications) when AR effect was closer to zero, or the amount of measurement error was large. In practice, hitting such a lower bound for the measurement error variance may erroneously suggest to researchers that the model is overly complex, and that there is no notable measurement error present in the data, which is problematic.

Note that while 100 observations may be small for estimation purposes, it is quite a large number of repeated measures to collect in practice. In psychological research using intensive longitudinal data, we usually see no more than about 120 observations per person (to illustrate, 120 observations would arise from about 4 months of daily measurements, or for more intense 2 weeks regime, measuring someone 9 times a day). Fortunately, the Bayesian AR(1)+WN model provides a good option even for such small sample sizes. Still, the models that incorporate measurement error need more observations to give as precise estimates as the basic AR(1) model, which has relatively small credible/confidence intervals (although this is precision around a wrong estimate when there actually is measurement error present in the data). Therefore, it seems good practice to take potential measurement error into account in the design of the study, thus collecting more repeated measures in order to compensate for any potential measurement error that has to be filtered out later. Expectedly, and as is shown in the simulation study, this becomes especially important when the proportion of measurement error variance is relatively large, or when the AR effects are (expected to be) relatively small. One option to improve the estimates may be to use (weakly) informative prior specifications based on previous research, or expert knowledge. However, prior information on the model parameters may currently prove difficult to obtain, given that studies that estimate measurement error or take it into account are very rare, and that the model parameters differ from person to person, and from variable to variable. Another option could be to extend the AR+WN model to a multilevel model, assuming a common distribution for the parameters of multiple individuals, and allowing the model parameters to vary across persons. By making use of this hierarchical structure that can take similarities between persons into account, a relatively low number of time points may be compensated for to some extent by a large number of participants, which may be easier to obtain (for examples of the

multilevel AR(1) model, see Rovine and Walls, 2006; Lodewyckx et al., 2011; De Haan-Rietdijk et al., 2014).

The reader may wonder how one may determine if there is, or isn't, measurement error present in the data. One way to do this is to use information criteria to compare the AR(1) model with the ARMA(1,1) or AR(1)+WN model. Although a thorough study of model selection is beyond the scope of the current paper, we provide some preliminary evaluations of the model selection performance of the AIC, BIC, and DIC, in Supplementary Materials. We find that these criteria frequently incorrectly selects the simpler AR(1) model over the (true) AR(1)+WN model and ARMA(1,1) model, so that these criteria seem inappropriate for selecting between the AR(1) and the ARMA(1,1) model or the AR(1)+WN model in this context. Selecting between an AR(1)+WN model and an ARMA(1,1) model will also be problematic using standard information criteria, because the AR(1)+WN model may be considered a restricted (simpler) version of the ARMA(1,1) model (see Equation 8), while they have the same number of parameters, and thus the same penalty for complexity for many fit criteria. In that sense, when they have equal fit, the AR(1)+WN model may be preferred because it is the simpler model, but if this is not the case, it becomes more complicated to choose between the two. Directions for future research therefore are to establish information criteria for selecting between the AR(1)+WN model and the AR(1) and ARMA(1,1) model, perhaps using information criteria or Bayes factors developed for restricted parameters (c.f., Dudley and Haughton, 1997; Klugkist and Hoijtink, 2007; Kuiper et al., 2012). Although model selection using information criteria may prove complicated, it is important to note that the estimates for ϕ in the AR(1)+WN models seem to be reasonably accurate, even when there is no measurement error present in the data. Combined with the intuition that most psychological measurements will contain at least some measurement error, fitting the model that incorporates measurement error seems a relatively "safe bet."

Another interesting topic for future work is how measurement error affects estimates of the effects variables have on each other over time, that is, the cross-lagged effects. This may be especially relevant for individual network models of psychological processes (Schmittmann et al., 2013). For example, in a network model for an individual diagnosed with a depressive disorder, the depression symptoms constitute the nodes in the network, and the AR and cross-lagged effects between the symptoms constitute the connections in this network (Borsboom and Cramer, 2013; Bringmann et al., 2013). It would be interesting to investigate to what extent measurement error in each variable affects the resulting network.

Finally, while incorporating measurement error into time series models is likely to decrease distortions as a result of ignoring measurement error to the parameter estimates, we emphasize that it is not a cure-all. Even in the models that incorporate measurement errors, the AR parameters may be slightly under- or over-estimated, because measurement error variance and innovation variance are not completely discernible from each other. The more measurement error present in the data, the more difficult it will be to pick up any effects. Therefore, there is still a strong argument for preventing measurement errors in the first place. One option to potentially improve the measurements is to use multiple indicators to measure the relevant construct. However, in a intensive longitudinal data setting, using multiple items for each variable would strongly increase the burden on the participant, who would have to repeatedly fill out all these questions. What remains are classical ways of preventing measurement error: Improving the respective measurement instruments, the circumstances under which participants are measured, and explicitly measuring and modeling potential sources of measurement error. Still, any remaining measurement error that could not be prevented, should be taken into account in the respective model. That is, prevention is better than cure—but a cure is better than ignoring the issue.

Acknowledgments

This study was supported by the Netherlands Organization for Scientific Research (NWO; VIDI Grant 452-10-007).

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2015.01038

References

Adolf, J., Schuurman, N. K., Borkenau, P., Borsboom, D., and Dolan, C. V. (2015). Measurement invariance within and between subjects: a distinct problem in testing the equivalence of intra- and inter-individual model structures. *Front. Psychol.* 5:883. doi: 10.3389/fpsyg.2014.00883

Borsboom, D., and Cramer, A. (2013). Network analysis: an integrative approach to the structure of psychopathology. *Annu. Rev. Clin. Psychol.* 9, 91–121. doi: 10.1146/annurev-clinpsy-050212-185608

Borsboom, D., Mellenbergh, G., and van Heerden, J. (2003). The theoretical status of latent variables. *Psychol. Rev.* 110, 203–219. doi: 10.1037/0033-295X.110.2.203

Bringmann, L., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., et al. (2013). A network approach to psychopathology: new insights into clinical longitudinal data. *PLoS ONE* 8:e60188. doi: 10.1371/journal.pone.0060188

Brooks, S. P., and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Stat.* 8, 434–455.

Casella, G., and George, E. I. (1992). Explaining the gibbs sampler. *Am. Stat.* 46, 167–174.

Chanda, K. C. (1996). Asymptotic properties of estimators for autoregressive models with errors in variables. *Ann. Stat.* 24, 423–430. doi: 10.1214/aos/1033066218

Chatfield, C. (2004). *The Analysis of Time Series: An Introduction.* Boca Raton, FL: Chapman and Hall; CRC.

Chong, T. T., Liew, V., Zhang, Y., and Wong, C. L. (2006). Estimation of the autoregressive order in the presence of measurement errors. *Econ. Bull.* 3, 1–10.

Cohn, J. F., and Tronick, E. (1989). Specificity of infants' response to mothers' affective behavior. *Adolesc. Psychiatry* 28, 242–248. doi: 10.1097/00004583-198903000-00016

Costa, M., and Alpuim, T. (2010). Parameter estimation of state space models for univariate observations. *J. Stat. Plan. Inference* 140, 1889–1902. doi: 10.1016/j.jspi.2010.01.036

De Haan-Rietdijk, S., Gottman, J. M., Bergeman, C. S., and Hamaker, E. L. (2014). Get over it! a multilevel threshold autoregressive model for state-dependent affect regulation. *Psychometrika*. doi: 10.1007/s11336-014-9417-x. [Epub ahead of print].

Dedecker, J., Samson, A., and Taupin, M. (2011). Estimation in autoregressive model with measurement error. *ESAIM Probab. Stat.* 18, 277–307. doi: 10.1051/ps/2013037

Deistler, M. (1986). "Linear dynamic errors-in-variables models," in *Contributions to Stochastics*, ed W. Sendler (Heidelberg: Physica-Verlag), 23–39.

Dudley, R. M., and Haughton, D. (1997). Information criteria for multiple data sets and restricted parameters. *Stat. Sin.* 7, 265–284.

Dunson, D. B. (2001). Commentary: Practical advantages of bayesian analysis of epidemiologic data. *Am. J. Epidemiol.* 153, 1222–1226. doi: 10.1093/aje/153.12.1222

Ferrer, E., Steele, J. S., and Hsieh, F. (2012). Analyzing the dynamics of affective dyadic interactions using patterns of intra- and interindividual variability. *Multivariate Behav. Res.* 47, 136–171. doi: 10.1080/00273171.2012.640605

Geller, E. S., and Pitz, G. F. (1968). Confidence and decision speed in the revision of opinion. *Organ. Behav. Hum. Perform.* 3, 190–201. doi: 10.1016/0030-5073(68)90005-6

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2003). *Bayesian Data Analysis, 2nd Edn.* Boca Raton, FL: Chapman and Hall;CRC.

Gelman, A., and Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Stat. Sci.* 7, 457–511. doi: 10.1214/ss/1177011136

Gilden, D. L. (2001). Cognitive emissions of 1/f noise. *Psychol. Rev.* 108, 33–56. doi: 10.1037/0033-295X.108.1.33

Goodwin, W. (1971). Resistance to change. *Am. Behav. Sci.* 14. 745–766. doi: 10.1177/000276427101400507

Granger, C. W. J., and Morris, M. J. (1976). Time series modelling and interpretation. *J. R. Stat. Soc. Ser. A* 139, 246–257. doi: 10.2307/2345178

Hamaker, E. (2012). "Why researchers should think "within-person": a paradigmatic rationale," in *Handbook of Research Methods for Studying Daily Life*, eds M. Mehl and T. Conner (New York, NY: Guilford Publications), 43–61.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton, NJ: Princeton University Press.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.

Hoijtink, H., Klugkist, I., and Boelen, P. (2008). *Bayesian Evaluation of Informative Hypotheses*. New York, NY: Springer.

Kievit, R., Romeijn, J., Waldorp, L., Wicherts, J., Scholte, H., and Borsboom, D. (2011). Mind the gap: a psychometric approach to the reduction problem. *Psychol. Inq.* 22, 67–87. doi: 10.1080/1047840X.2011.550181

Kim, C.-J., and Nelson, C. (1999). *State–Space Models with Regime Switching*. Cambridge, MA: The MIT Press.

Kirkham, N. Z., Cruess, L., and Diamond, A. (2003). Helping children apply their knowledge to their behavior on a dimension-switching task. *Dev. Sci.* 5, 449–476. doi: 10.1111/1467-7687.00300

Klugkist, I., and Hoijtink, H. (2007). The bayes factor for inequality and about equality constrained models. *Comput. Stat. Data Anal.* 51, 6367–6379. doi: 10.1016/j.csda.2007.01.024

Koval, P., Kuppens, P., Allen, N. B., and Sheeber, L. (2012). Getting stuck in depression: the roles of rumination and emotional inertia. *Cogn. Emot.* 26, 1412–1427. doi: 10.1080/02699931.2012.667392

Kuiper, R., Hoijtink, H., and Silvapulle, M. (2012). Generalization of the order-restricted information criterion for multivariate normal linear models. *J. Stat. Plann. Inf.* 142, 2454–2463. doi: 10.1016/j.jspi.2012.03.007

Kuppens, P., Allen, N. B., and Sheeber, L. B. (2010). Emotional inertia and psychological maladjustment. *Psychol. Sci.* 21, 984–991. doi: 10.1177/0956797610372634

Lee, M. D., and Wagenmakers, E.-J. (2005). Bayesian statistical inference in psychology: comment on trafimow (2003). *Psychol. Rev.* 112, 662–668. doi: 10.1037/0033-295X.112.3.662

Lodewyckx, T., Tuerlinckx, F., Kuppens, P., Allen, N., and Sheeber, L. (2011). A hierarchical state space approach to affective dynamics. *J. Math. Psychol.* 55, 68–83. doi: 10.1016/j.jmp.2010.08.004

Luethi, D., Erb, P., and Otziger, S. (2010). *FKF: Fast Kalman Filter*. R Package Version 0.1.1.

Madhyastha, T., Hamaker, E., and Gottman, J. (2011). Investigating spousal influence using moment-to-moment affect data from marital conflict. *J. Fam. Psychol.* 25, 292–300. doi: 10.1037/a0023028

Moberly, N., and Watkins, E. (2008). Ruminative self–focus and negative affect: an experience sampling study. *J. Abnorm. Psychol.* 117, 314–323. doi: 10.1037/0021-843X.117.2.314

Molenaar, P. (2004). A manifesto on psychology as idiographic science: bringing the person back into scientific psychology, this time forever. *Measurement* 2, 201–218. doi: 10.1207/s15366359mea0204.1

Nezlek, J., and Allen, M. (2006). Social support as a moderator of day–to–day relationships between daily negative events and daily psychological well–being. *Eur. J. Pers.* 20, 53–68. doi: 10.1002/per.566

Nezlek, J. and Gable, S. (2001). Depression as a moderator of relationships between positive daily events and day–to–day psychological adjustment. *Pers. Soc. Psychol. Bull.* 27, 1692–1704. doi: 10.1177/01461672012712012

Patriota, A. G., Sato, J. R., and Blas Achic, B. G. (2010). Vector autoregressive models with measurement errors for testing granger causality. *Stat. Methodol.* 7, 478–497. doi: 10.1016/j.stamet.2010.02.001

Plummer, M. (2003). *Jags: A Program for Analysis of Bayesian Graphical Models using Gibbs Sampling.*

Plummer, M., Stukalov, A., and Plummer, M. M. (2014). *Package Rjags: Update*.

R Development Core Team. (2012). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rietbergen, C., Groenwold, R. H., Hoijtink, H. J., Moons, K. G., and Klugkist, I. (2014). Expert elicitation of study weights for bayesian analysis and meta-analysis. *J. Mixed Methods Res.* doi: 10.1177/1558689814553850. (in press).

Rietbergen, C., Klugkist, I., Janssen, K. J., Moons, K. G., and Hoijtink, H. J. (2011). Incorporation of historical data in the analysis of randomized therapeutic trials. *Contemp. Clin. Trials* 32, 848–855. doi: 10.1016/j.cct.2011.06.002

Rovine, M., and Walls, T. (2006). "A multilevel autoregressive model to describe interindividual differences in the stability of a process," in *Models for Intensive Longitudinal Data*, eds J. Schafer and T. Walls (New York, NY: Oxford), 124–147.

Schmittmann, V., Cramer, A., Waldorp, L., Epskamp, S., Kievit, R., and Borsboom, D. (2013). Deconstructing the construct: a network perspective on psychological phenomena. *New Ideas Psychol.* 31, 43–53. doi: 10.1016/j.newideapsych.2011.02.007

Staudenmayer, J., and Buonaccorsi, J. P. (2005). Measurement error in linear autoregressive models. *J. Am. Stat. Assoc.* 100, 841–852. doi: 10.1198/016214504000001871

Suls, J., Green, P., and Hillis, S. (1998). Emotional reactivity to everyday problems, affective inertia, and neuroticism. *Pers. Soc. Psychol. Bull.* 24, 127–136. doi: 10.1177/0146167298242002

Swamy, P. A. V. B., Chang, I., Mehta, J. S., and Tavlas, G. S. (2003). Correcting for omitted-variable and measurement-error bias in autoregressive model estimation with panel data. *Comput. Econ.* 22, 225–253. doi: 10.1023/A:1026189916020

Wagenmakers, E., Farrell, S., and Racliff, R. (2004). Estimation and interpretation of 1/f noise in human cognition. *Psychon. Bull. Rev.* 11, 579–615. doi: 10.3758/BF03196615

Wang, L., Hamaker, E., and Bergeman, C. (2012). Investigating inter-individual differences in short-term intra-individual variability. *Psychol. Methods* 17, 567–581. doi: 10.1037/a0029317

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Schuurman, Houtveen and Hamaker. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in PSYCHOLOGY

OPINION ARTICLE
published: 28 November 2014
doi: 10.3389/fpsyg.2014.01356

The implication of the coefficient of centrality for assessing the meaning of the mean

David Trafimow*

Department of Psychology, New Mexico State University, Las Cruces, NM, USA
**Correspondence: dtrafimo@nmsu.edu*

Edited by:
Craig Speelman, Edith Cowan University, Australia

Reviewed by:
Michael Smithson, Australian National University, Australia

Keywords: coefficient of variation, coefficient of centrality, meaning, standard deviation, meaning of mean

The coefficient of variation (C_V) is an important and underused statistic that implies that the standard deviation has different meanings depending on the mean (Fisher, 1925; Yates, 1951; Yadav et al., 2013; Trafimow, 2014) and is computed by dividing the standard deviation by the mean $\left(C_V = \frac{\sigma}{\mu}\right)$. To gain a feel for how means impact the meanings of standard deviations, imagine that there are two classes and that the standard deviation of the final exam is 5 points in both of them but that the mean is 25 in Class 1 and 75 in Class 2. It follows that the coefficient of variation is 0.200 in Class 1 whereas it is 0.067 in Class 2. Thus, relative to the mean the standard deviation is thrice the size in Class 1 as in Class 2. Consequently, the actual value of the standard deviation is the same in both classes but its meaning is very different.

In keeping with this Special Issue on means, my goal is not to discuss how means influence the meanings of standard deviations (see Trafimow, 2014 for this discussion). Rather, it is to show that the coefficient of variation is a two edged sword so that if means modify the meanings of standard deviations, the reverse also is so. Standard deviations influence the meanings of means. An easy demonstration involves defining a new variable, termed the *coefficient of centrality* (C_C), which is the reciprocal of the coefficient of variation and is given as Equation (1) below.

$$C_C = \frac{1}{C_V} = \frac{1}{\frac{\mu}{\sigma}} = \frac{\mu}{\sigma} \qquad (1)$$

To gain a preliminary idea of how the coefficient of centrality works, suppose Company A makes pies with a mean of 100 per day and distributes them to a store that, on average, sells 100 of the pies per day. Attending only to means, life seems good because Company A is producing exactly the number of pies that maximizes profit but avoids the perils of overproduction. But now consider the standard deviation. Suppose that the standard deviation is 15 so that there is an approximately 16% chance that on any particular day, the factory will be short by 15 or more pies and a 16% chance that the factory will be long by 15 or more pies. Now, imagine that the company makes an innovation that reduces the standard deviation from 15 to 5. From a mean-centric point of view, the innovation might seem irrelevant because the mean remains at 100 pies per day. But the coefficient of centrality suggests otherwise as the innovation causes an increase in the coefficient from 6.67 to 20. Now the probability of underproduction by 15 or more pies decreases from 16% to 0.15% and the probability of overproduction by 15 or more pies decreases similarly. Clearly, the same mean value of 100 pies per day has different implications for underproduction and overproduction depending on the standard deviation.

Let us now consider Company B that also makes pies. This company produces a mean of 105 pies per day, with a standard deviation of 15 pies, even though their outlet is only willing to buy 100 pies per day. A possible reason for overproducing is that it is much worse to anger the customer by not having enough pies than to overproduce. Assuming this reason is valid, if we take Company A before its innovation when it also had a standard deviation of 15 but a mean of 100, it is obvious that the mean of Company B is higher than the mean of Company A, and Company A is therefore more at risk of angering its customer base. But let us now compare the means of the two companies after the Company A innovation reduced its standard deviation to 5. Which mean is larger? It depends on what we mean by "larger." At first blush, the mean is 105 for Company B and 100 for Company A and so the mean is larger for Company B than for Company A. On the other hand, consider the coefficients of centrality; these are 7 for Company B and 20 for Company A and suggest the opposite conclusion. Which conclusion is correct? It depends on the goal. If the goal were simply to maximize pie production over a period of time, then a mean of 105 is superior to a mean of 100. But if the goal is to avoid dramatic underproduction, then the latter conclusion is correct; Company B (despite the mean of 105) will have more days of dramatic underproduction than will Company A (despite the mean of 100). The coefficient of centrality demonstrates that means have different meanings depending on standard deviations.

Consider a more basic example. One professor teaches an undergraduate class on abnormal psychology and another professor teaches an undergraduate class on cognitive psychology where the mean scores on the final exam are 75% and 65%, respectively. Is the abnormal psychology class better than the cognitive psychology class? Suppose that the standard deviations are 25% and 5% so that the coefficients of centrality are 3 and 13, respectively.

Relative to the standard deviations, the cognitive psychology class mean well exceeds the abnormal psychology class mean, which suggests exactly the opposite conclusion. Of course, there are many other factors that could be at play but the coefficient of centrality suggests that it can be a mistake to consider the means without also considering the standard deviations.

In light of this example, it is worth mulling over advanced statistical literatures pertaining to standard deviation weighted analysis of variance (Kulinskaya et al., 2003) and weighted least squares linear regression (Strutz, 2010). These analyses result in means that are weighted by standard deviations to take differing standard deviations (heteroscedasticity) into account. Given the availabilities of the proposed coefficient of centrality and these advanced methods, it is difficult to justify researchers routinely failing to consider standard deviations when interpreting their means in future research, regardless of how complicated the data happen to be.

REFERENCES

Fisher, R. A. (1925). *Statistical Methods for Research Workers on the Development of the Science of Statistics*. Edinburgh: Oliver and Boyd.

Kulinskaya, E., Staudte, R. G., and Gao, H. (2003). Power approximations in testing for unequal means in a one-way ANOVA weighted for unequal variances. *Commun. Stat. Theory Methods* 32, 2353–2371. doi: 10.1081/STA-120025383

Strutz, T. (2010). *Data Fitting and Uncertainty (A Practical Introduction to Weighted Least Squares and Beyond)*. Wiesbaden: Vieweg + Teubner.

Trafimow, D. (2014). On teaching about the coefficient of variation in introductory statistics courses. *Teach. Stat.* 36, 81–82. doi: 10.1111/test.12042

Yadav, R., Upadhyaya, L. N., Singh, H. P., and Chatterjee, S. (2013). A general procedure of estimating the population variance when coefficient of variation of an auxiliary variable is known in sample surveys. *Qual. Quant. Int. J. Methodol.* 47, 2331–2339. doi: 10.1007/s11135-012-9659-6

Yates, F. (1951). The influence of statistical methods for research workers on the development of the science of statistics. *J. Am. Stat. Assoc.* 46, 19–34. doi: 10.2307/2280090

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 24 September 2014; accepted: 07 November 2014; published online: 28 November 2014.
*Citation: Trafimow D (2014) The implication of the coefficient of centrality for assessing the meaning of the mean. Front. Psychol. **5**:1356. doi: 10.3389/fpsyg.2014.01356*
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology.
Copyright © 2014 Trafimow. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read, for greatest visibility

COLLABORATIVE PEER-REVIEW
Designed to be rigorous – yet also collaborative, fair and constructive

FAST PUBLICATION
Average 85 days from submission to publication (across all journals)

COPYRIGHT TO AUTHORS
No limit to article distribution and re-use

TRANSPARENT
Editors and reviewers acknowledged by name on published articles

SUPPORT
By our Swiss-based editorial team

IMPACT METRICS
Advanced metrics track your article's impact

GLOBAL SPREAD
5'100'000+ monthly article views and downloads

LOOP RESEARCH NETWORK
Our network increases readership for your article

Frontiers
EPFL Innovation Park, Building I • 1015 Lausanne • Switzerland
Tel +41 21 510 17 00 • Fax +41 21 510 17 01 • info@frontiersin.org
www.frontiersin.org

Find us on