Overcoming experimental psychology's replicability crisis by studying humans at the intra- rather than inter-individual level

Etienne P. LeBel
Montclair State University

Denny Borsboom
University of Amsterdam

Fred Hasselman
Radboud University Nijmegen

Christoph Stahl
University of Cologne

Manuscript version: 4.3 [July 2014]

Word count: 6,152

Corresponding Author: etienne.lebel@gmail.com

Abstract

Several important initiatives and methodological reforms have recently been implemented to help address psychology's current replicability crisis. Though necessary, we contend that these positive methodological developments are insufficient to overcome the current crisis because of a more fundamental deficiency of experimental research in psychology: Experimental psychologists theorize at the intra-individual level but test their hypotheses at the inter-individual level. In this article, we argue that overcoming experimental psychology's crisis requires psychologists to investigate human beings at the intra-individual level rather than the inter-individual level**,** which is currently the dominant approach in psychology. A meta-theoretic change in how we think about psychological phenomena is required, not just methodological changes aimed at improving modal research practices. To support our argument, we review inter-individual versus intra-individual approaches via an example from social psychology and another from cognitive psychology. We carefully clarify important differences in what kinds of research questions can be answered using the two different approaches. We make a case that experimental psychologists ultimately need to examine psychological phenomena at the intra-individual level given likely participant treatment heterogeneity, which invalidates general results found at the inter-individual level. We introduce a general meta-theoretic framework that allows one to examine psychological phenomena at the intra-individual level to identify and understand the nature of participant heterogeneity, illustrating the approach via two empirical demonstrations and applied examples from different areas in psychology. We conclude by discussing implications of our approach and addressing possible misconceptions.

Overcoming experimental psychology's replicability problem by studying humans at the intra- rather than inter-individual level

Experimental psychology is currently facing a "crisis of confidence" in that a growing number of published findings cannot be replicated via independent replication (Pashler & Wagenmakers, 2012; Pashler & Harris, 2012; Simons, 2014; OSC 2012, 2013). Many initiatives have been launched to help the situation by improving research practices, including pre-registration (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012; Chambers, 2013; Wolfe, 2013), open data (Simons, 2014; Nosek, Spies, & Motyl, 2012), higher reporting standards (Eich, 2013; Simmons, Leif, & Simonsohn, 2011; LeBel et al., 2013), increased emphasis on estimation based on effect sizes and confidence intervals (Cumming, 2014), and publication of direct replications (Neuliep & Crandall, 1990; Rosenthal, 1997; Schimmack, 2012).

Such initiatives are necessary and will help because science requires rigorous methods, independent verification of analyses, and independent corroboration of reported findings to avoid errors and ensure cumulative knowledge development (Feynman, 1974; Popper, 1934/1992). Though necessary, we contend that these positive developments are unfortunately insufficient to overcome experimental psychology's current crisis because experimental psychologists ultimately want to know about intra-individual psychological processes but investigate psychological phenomena almost exclusively at the inter-individual level. We assert that experimental psychology's real crisis is that experimental psychologist theorize at the intra-individual level but then test their hypotheses at the inter-individual level. In this article, we argue that **overcoming experimental psychology's current crisis requires psychologists to study human beings at the *intra-individual* rather than *inter-individual* level**, which is currently the dominant approach in experimental psychology. A meta-theoretic change in how we think about psychological phenomena is required, not solely methodological changes aimed at improving modal research practices.

Psychology's replicability crisis

There is growing consensus that experimental psychology is currently facing a "crisis of confidence" (Pashler & Wagenmakers, 2012; Pashler & Harris, 2012; Simons, 2014; OSC 2012, 2013), reflected in the fact that a growing number of published findings cannot be replicated when competent independent researchers execute high-powered replication attempts that duplicate the original methodology as closely as possible. For example, the Reproducibility Project, an open large-scale attempt at estimating the replicability of psychological science (OSC, 2012, 2013), were unable to replicate over 70% (20 out of 28, as of May 2014) of findings systematically selected from 2008 issues of *Psychological Science*, *Journal of Personality and Social Psychology*, and *Journal of Experimental Psychology: Learning, Memory, and Cognition.* In another large-scale meta-scientific investigation, about 60% (11 out of 27) of important findings from cognitive and social psychology could not be replicated (Nosek & Lakens, 2014). Furthermore, there is a growing list of (prominent) findings in all areas of psychology that have not held up to independent replication attempts, including findings from **cognitive psychology** (retrieval-induced forgetting, Maslany & Campbell, 2013; eye movements on recall, Matzke et al., 2014; temporal judgments, Matthews, 2012; protection effect, Wolferen, Inbar, &

Zeelenberg, 2013; mental simulation, Zwaan & Pecher, 2012; Mozart effect, Steele, Bass, & Crook, 1999), **developmental psychology** (synesthetic cross-modality correspondence, Lewkowicz & Minar, in press), **neurophysiology** (vestibular stimulation, Lenggenhager, Hilti, Palla, Macauda, & Brugger, 2014), **industrial/organizational psychology** (utility biasing effect on selection procedures, Carson, Becker, & Henderson, 1998), **positive psychology** (weather effects on life satisfaction, Schmiedeberg & Schroder, 2014), **political psychology** (self-prophecy effect on voting, Smith, Gerber, & Orlich, 2003; status-legitimacy hypothesis, Brandt, 2013), **moral psychology** ("Macbeth effect", Earp, Everett, Madva, & Hamlin, 2013), **educational psychology** (stereotype threat on math performance, Ganley et al., 2013; color influence on exam performance, Tal, Akers, & Hodge, 2008), **evolutionary psychology** (fertility on face preferences, Harris, 2011; ovulation on men's testosterone, Roney & Simmons, 2012; sex differences in infidelity distress, IJzerman et al., 2014**), judgment & decision making** (unconscious thought effects, Nieuwenstein & van Rijn, 2012), and the so-called area of **behavioral priming/embodiment** (Johnson, Cheung, & Donnellan, 2014; Donnellan, Lucas, & Cesario, in press; Lynott et al., 2014; Doyen, Klein, Pichon, & Cleeremans, 2012; Harris, Coburn, Rohrer, & Pashler, 2013; McCarthy, in press; Pashler, Rohrer, & Harris, in press; Pecher, 2014; Rohrer, Pashler, & Harris, 2014; Steele et al., 2013; LeBel & Campbell, 2013; LeBel & Wilbur, 2014; Madurski & LeBel, 2014). Taken together, these observations lead to the conclusion that there experimental psychology is currently facing a general crisis of confidence rather than isolated cases of doubts regarding specific areas of experimental psychology.

New initiatives aimed at improving research practices

Several new initiatives have been launched to improve research practices in order to increase the reliability of findings in psychology. For instance, there have been major developments regarding requiring higher reporting standards at several prominent psychology journals (Eich, 2013; Simmons, Leif, & Simonsohn, 2011, 2012; LeBel et al., 2013). At such journals (e.g., *Psychological Science, Memory & Cognition, Learning & Behavior, Attention, Perception, & Psychophysics, Psychonomic Bulletin & Review, Personality and Social Psychology Bulletin, Cognitive, Affective, & Behavioral Neuroscience*), authors submitting a manuscript must now acknowledge that they have disclosed basic methodological details critical for the accurate evaluation and interpretation of reported findings such as fully disclosing all excluded observations, all tested experimental conditions, all assessed outcome measures, and their data collection termination rule.

There is also a significant push to incentivize "open data", the public posting of the raw data underlying studies reported in a published article (Nosek, Spies, & Motyl, 2012; Simons, 2014). For instance, at *Psychological Science,* authors who make their data publicly available earn an open data badge that is prominently displayed alongside their published article and the new *Journal of Open Psychology Data* now publish data papers that feature publicly posted data sets (Wicherts, 2013). Such open data practices not only facilitate independent verification of analyses and results so crucial to identifying errors and other inaccuracies, but substantially facilitate the execution of meta-analyses and re-analyses from different theoretical perspectives, which can lead to new insights and accelerate knowledge development.

Another development is an increased emphasis on “estimation” based on effect size and confidence intervals rather than exclusive focus on dichotomous $p < .05$ thinking (Cumming, 2014). From this perspective, the goal is to iteratively execute increasingly rigorous studies (e.g., by using larger samples sizes, stronger measures), to arrive at a more precise estimate of the effect size of a certain psychological phenomenon (Simons, 2014; see also LeBel 2014, Simonsohn, 2014).

In addition, several journals (e.g., *Cortex, Perspectives on Psychological Science*, *Attention, Perception, & Psychophysics,*) now offer pre-registered publication options whereby authors submit a study proposal that pre-specifies the methodology and analytical approaches to be used to answer a certain research question (Chambers, 2013; Wolfe, 2013). Proposals are evaluated on the soundness of the methodology and theoretical importance of the research question. Once accepted, the proposed study is executed and the article is published regardless of the results. This way, questionable research practices – either by researchers or editors – which can grossly mischaracterize evidence are avoided which will lead to a more accurate published literature (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012).

Finally, yet another development that we find particularly exciting, is the growing practice of prominent journals to publish independent direct replication results, including replication results inconsistent with those originally published by the journal (e.g., *Psychological Science, Psychonomic, Bulletin, & Review, Journal of Experimental Social Psychology*). Though calls for the publication of replication results have been made for decades (e.g., Neuliep & Crandall, 1990; Rosenthal, 1997), the actual practice of prominent journals of systematically publishing replication results is unprecedented and has immense potential to improve the reliability of findings in psychology. Such practice not only directly incentivizes researchers to actually execute independent replications so crucial to corroborating original results and ensuring a cumulative knowledge base (Feynman, 1974; Popper, 1934/1992), but also should significantly reduce the tendency for researchers to publish unexpected and/or tenuous results (Schimmack, 2012).

We strongly believe that such varied initiatives will substantially help improve psychological science because any scientific discipline requires extreme rigor in its methods, independent verification of analyses, and independent corroboration of reported findings to avoid errors and ensure cumulative knowledge development (Feynman, 1974; Popper, 1934/1992). However, we contend that these positive developments – though necessary -- will unfortunately be insufficient to overcome experimental psychology’s current crisis because of a more fundamental fact: **experimental psychologists really want to know about intra-individual psychological processes but almost exclusively study psychological phenomena at the inter-individual level.** In other words, we maintain that even if *all* psychologists fully adopted *all* of these new research practices and recommendations, psychology would still have a crisis on their hands because of a fundamental disconnect between what psychologists want to know (i.e., intra-individual processes) and the modal approach used (i.e., inter-individual level approach).

Psychological phenomena need to be examined at the intra-individual level

In this article, **we argue that overcoming experimental psychology's crisis requires psychologists to study human beings at the intra-individual level rather than the inter-individual level,** which is currently the dominant approach in psychology. In other words, a meta-theoretic change in how we think about psychological phenomena is required, not just methodological changes aimed at improving modal research practices. To support our argument, we will begin by reviewing inter-individual versus intra-individual approaches via two example findings (one from social, one from cognitive). We then carefully clarify important differences in what kinds of research questions can be answered using the two different approaches. We then make a case that experimental psychologists ultimately need to examine psychological phenomena at the intra-individual level given likely participant treatment heterogeneity, which invalidates general results found at the inter-individual level. We then introduce a general meta-theoretic framework that allows one to examine psychological phenomena at the intra-individual level to identify and understand the nature of participant heterogeneity, illustrating the approach via two empirical demonstrations and applied examples from different areas in psychology. We then discuss implications of our approach and address possible misconceptions and concerns of our approach.

Two disciplines of psychology: Inter vs. Intra

The modal inter-individual level approach. The modal research approach in psychology – that is, the most frequently used approach – examines psychological phenomena at the inter-individual level of analysis, which involves examining mean differences or co-variability between constructs *across* individuals. For example, such an approach has traditionally been used in studies in the area of persuasion to examine the effects of consensus influence strategies, whereby individuals' beliefs and behaviors are influenced by being aware that many other individuals exhibit such behaviors. For example, in a typical experiment, participants would be randomly assigned to either a consensus (e.g., "Over a million copies sold!") or control (e.g., "A random selection from our inventory") condition and subsequently all participants would be asked to report their purchase intentions for the same set of consumer products (Cialdini, 2001). Mean purchase intention scores would then be compared across consensus and control (between-subject) conditions. As depicted in Figure 1 (left panel), a typical data pattern found in such studies is that mean purchase intention scores in the consensus condition are statistically significantly higher than mean purchase intention scores in the control condition, which is taken as evidence that a consensus influence strategy boosts purchase intentions (Cialdini, 2004). It is important to explicitly point out, however, that such evidence involves averaging *across* individual subjects (as can be seen in Figure 1, each subject is only in one of the conditions).

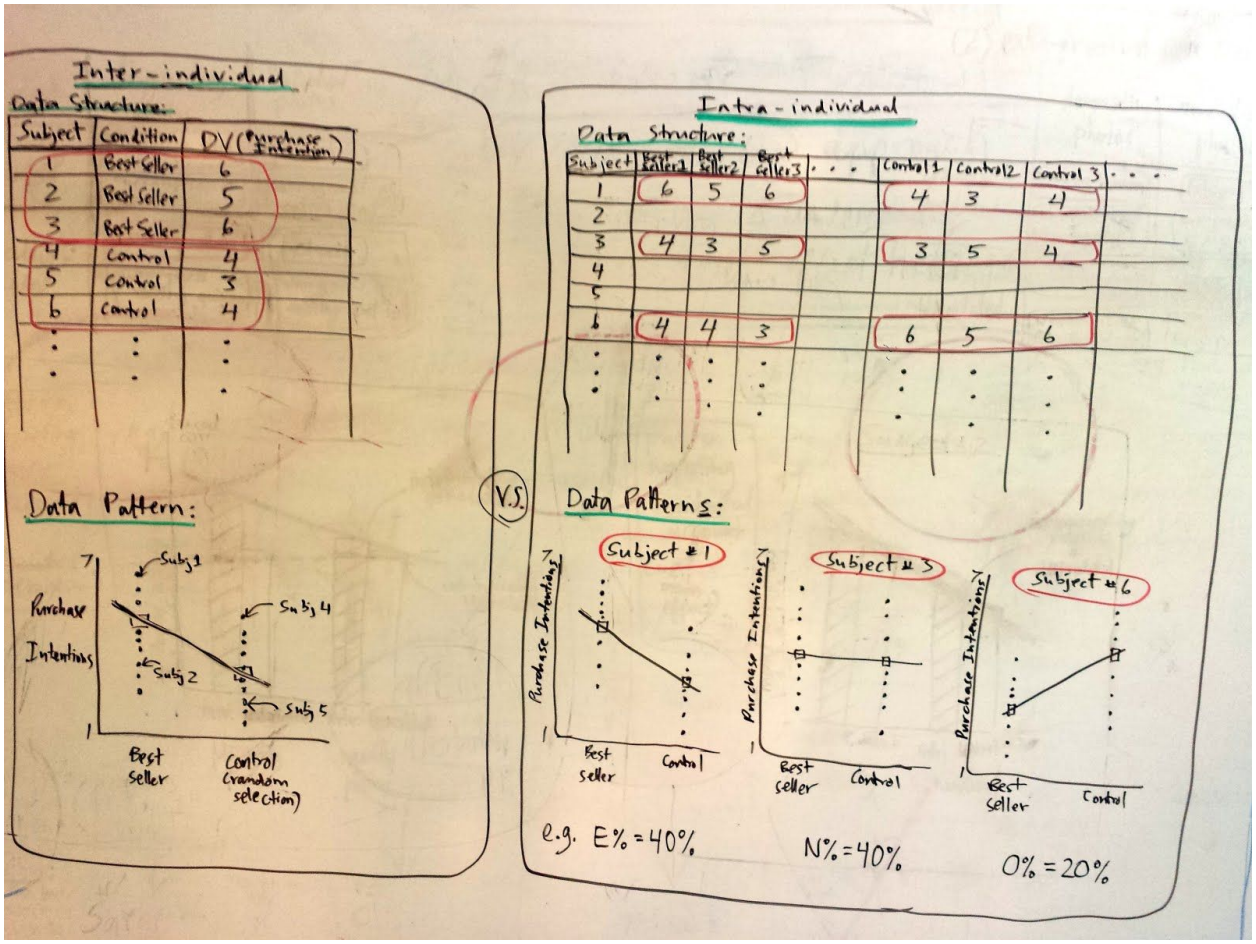

Figure 1. Data structure and possible data pattern(s) for inter-individual (left) and intra-individual (right) designs.

Though the consensus influence strategy effects (of the sort just described) have been widely documented in different contexts and domains (Cialdini, 2001, 2004), there have also been reports of “difficulties in replications” (Kaptein & van Halteren, 2011, p. 1175) and inconsistent results for closely related effects of argument strength on persuasion (Johnson & Eagly, 1989).

A second example of the modal inter-individual approach, drawn from cognitive psychology, using a within-subjects (rather than between-subjects) design, involves a perceptual simulation effect (Estes, Verges, & Barsalou, 2008). Across three studies, Estes et al. demonstrated that “words denoting objects that typically occur in high places (e.g., hat, cloud) hindered identification of targets appearing at the top of the display, whereas words denoting objects that typically occur in low places (e.g., boot, puddle) hindered identification of targets at the bottom” (p. 96), presumably because individuals were perceptually simulating the denoted object which interfered with the perception of targets. This effect was argued to shed important new light on how language affects attention and perception. Again, it is important to explicitly point out that the effect in question here was that mean reaction times (RTs) to identify targets – averaged across individuals – were longer when they appeared in the typical location of the object denoted by the preceding word than when they appeared in the opposite (atypical)

location. Though this effect was based on a large body of behavioral and neural evidence supporting grounded cognition (i.e., Barsalou, 2008, now cited over 2,150 times), subsequent high-powered independent replication attempts have failed to corroborate the perceptual simulation effect (Renkewitz & Muller, 2012) and replication difficulties have also emerged for other related grounded cognition effects (Zwaan & Pecher, 2012; see also LeBel & Campbell, 2013; LeBel & Wilbur, 2014). Taken together, these two examples demonstrate that plausible psychological phenomena based on prior research and sound theoretical reasoning – but more importantly documented using the modal inter-individual approach – have subsequently experienced replication difficulties.

The intra-individual level. As can be seen in Figure 1, examining the question of how influence strategies affects persuasion at the intra-individual level requires using a completely different experimental design. In such a design, each individual subject would be repeatedly exposed to all levels of the independent variable (e.g., consensus and control conditions) and the dependent variable would be repeatedly assessed (e.g., purchase intentions). In this way, we can determine, for each individual subject, whether an independent variable had a statistically significant influence on a dependent variable and if so, in which direction (e.g., an independent variable increased or decreased a certain behavior).

Importantly, results found at the inter-individual level may not generalize to the intra-individual level. Indeed, it is completely possible that a phenomenon at the inter-individual doesn't emerge for *any* of the individuals in one's sample. Indeed, below we review theorizing that indicates that only under very strict conditions of ergodicity, unlikely to be found for psychological phenomena, do results at the inter-individual level generalizes to the intra-individual level. We also review theoretical reasons to expect heterogeneity across individuals, whereby different individuals are influenced differently by the same experimental manipulation, and cite several documented examples of such heterogeneity. Given these considerations, we argue that experimental psychologists ultimately need to examine psychological phenomena at the intra-individual level to really understand the cognitive and affective processes underlying human behavior. Then, and only then, will we be able to overcome the current crisis in experimental psychology.

Many reasons to expect heterogeneity

A simple way to begin this discussion is to imagine a particular experimental manipulation, whether in the form of exposing participants to words, photos, specific instructions or a particular situation. If there's any *plausible* possibility that the same manipulation can be interpreted, perceived, or experienced differently by different individuals in one's sample, then averaging across such individuals will likely lead to small effects that will be prohibitively difficult to detect. Coming back to our two examples from above, replication difficulties of the scarcity effect (e.g., Kaptein & Eckles, 2012; Jung & Kellaris, 2004), may stem from heterogeneity in the scarcity effect across individuals whereby it emerges only for certain individuals, but not in others. This could be due to various factors. Perhaps different participants interpreted the perceived *cause(s)* of the product shortage differently (i.e., perceived scarcity, a factor Worchel et al., 1975 attempted to examine in their original investigation). For example, particular participants perceiving the product shortage due to stocking problems would not be expected to

show a scarcity effect (or to show a smaller effect) whereas particular participants perceiving the shortage due to high consumer demand *would* be expected to show a scarcity effect. Alternatively, perhaps different participants had different levels of product familiarity, uncertainty avoidance, and/or need for cognitive closure, all factors implicated in the scarcity effect (Jung & Kellaris, 2004; e.g., smaller scarcity effects for individuals more familiar with a product).

Such heterogeneity issues are also relevant to the related consensus influence strategy effect, whereby individuals who have esoteric tastes may actually be *less* (rather than more) influenced by a consensus persuasive message (e.g., claiming a certain product is a best-seller). Indeed, there have also been replication difficulties of social-norms marketing effects used to curb undesirable and/or unhealthy behaviors (e.g., alcohol consumption, gambling, littering; Clapp, Lange, Russell, Shillington, & Voas, 2003; Granfield, 2005; Peeler, Far, Miller, & Brigham, 2000; Russell, Clapp, & DeJong, 2005; Werch et al., 2000), with several studies finding "boomerang effects" whereby social-norms marketing campaigns actually *increased* – rather than decreased – the targeted undesirable behaviors (e.g., Perkins, Haines, & Rice, 2005; Wechsler et al., 2003; Werch et al., 2000). Social-norms marketing campaigns are supposed to reduce problem behaviors by communicating via a descriptive norm message that a problem behavior occurs less often than most people think (e.g., most undergraduates only consume 4-5 alcoholic drinks in a sitting), whereby the descriptive norm acts as a magnet to guide individuals' behavior. For example, binge drinkers who typically consume 10 drinks per sitting will attempt to adjust their behavior closer to the 4-5 drinks descriptive norm, however, individuals who typically drink much less than the descriptive norm may actually end up drinking more, resulting in a boomerang effect. Hence, the same social-norm manipulation can influence different individuals differently because the stated information is interpreted differently by different individuals vis-à-vis personal characteristics they bring to the situation (e.g., their own personal drinking patterns).

The same logic applies for the replication difficulties involving the perceptual simulation effect from the cognitive psychology literature. The effect hinges upon the assumption that the object words will consistently activate the relevant spatial representations across individuals (e.g., COWBOY HAT will lead to perceptually simulating a hat in its typically relatively high position). This may not be the case, however, for particular individuals that physically perceive most of the objects atypically (e.g., "DOG PAW" may elicit the high position for dog owners/trainers who recently trained a dog to give "high fives") or due to the fact that particular individuals may simply not associate most of the words with those spatial/physical positions (e.g., exposure to "HOTEL LOBBY" may lead an individual to conjure up an image of the last hotel lobby they were in rather than a "low" position in space).

Indeed, extant theorizing more formally points to several different factors that can directly contribute to heterogeneity of psychological phenomena across individuals, including **culture**, **social class, geography, occupation**, and emotionally-significant **life experiences**. Paul Rozin (2001) – in the context of the risks of over-emulating the natural sciences – speaks to this point with respect to culture and social class:

"It is probably true that if you understand one eyeball, you will understand them all, but it is not at all true that if you understand one person, you will understand them all. In particular, people's lives, behavior, and

mental events are strongly influenced and shaped by the culture they are members of [and] by the structure of their society and their place in it." (p. 10)

In other words, culture and social class can influence individuals' behaviors and mental events in substantial ways. Consequently, individuals from different cultures and/or social class may be influenced by the same experimental manipulation differently, creating heterogeneity across individuals in the target psychological phenomenon being studied. We take "culture" here to refer to the shared knowledge, values, beliefs, and practices among a group of people living in geographical proximity (Atran, Medin, & Ross, 2005). It is important to point out, however, that culture and its influence can vary dramatically within the same country (e.g., culture of honor in southern vs. northern U.S., Nisbett & Cohen, 1996) and within culture (e.g., categorization processes among rural vs urban children, Medin & Atran, 2004).

Social class can also have powerful influences on the way a particular individual interprets or experiences information or stimuli embedded in an experimental manipulation, leading to heterogeneity across individuals. Bruner and Goodman's (1947) classic paper finding that 10 year olds from low social class over-estimated the size of coins moreso than 10 year olds from high social class (an effect also reported by Dawson, 1975 using a Hong Kong Chinese sample of children).

- Language/vocabulary variability (e.g., language influences categorization, e.g. Ji, Zhang, & Nisbett, 2004; language influences sensory processing in emotion perception, Barrett, Lindquist, & Gendron, 2007; language can influence perception more generally, Klemfuss, Prinzmetal, & Ivry, 2012)
- Occupation/Life experiences (e.g., significant life events on academic performance, Harris, 1973; life events on psychopathology, Paykel et al., 1969, Sarason, Johnson, & Siegel, 1978)
- Subjective interpretation/meaning of the situation (Peters, 2010; Mischel & Shoda, 1995; Mischel, 2007; check other references in Peters' paper)

Theorizing consistent with expecting heterogeneity:

- Different strategies to complete the same task (Luce, 1995, 1997)
- Perils of aggregating across individuals (e.g., Estes, 1956; Estes & Maddox, 2005)
- Questionable ergodicity assumption underlying modal research (e.g., Molenaar, 2004a, 2004b; Molenaar & Campbell, 2009; Borsboom et al., 2003; i.e., in the presence of between-person heterogeneity, average patterns (or inter-individual patterns in the case of correlational data) do *not* necessarily describe *any* of the individuals in a sample)

"An almost universal—but surprisingly silent—reliance on what may be called a uniformity-of-nature assumption in doing [aggregate level] analyses; the relation between mechanisms that operate at the level of the individual and models that explain variation [at the aggregate level] is often taken for granted, rather than investigated." (Borsboom et al., 2003, p. 215)

For many psychologists it would seem that *"the processes acting upon mental representations are construed to be essentially similar, not only across all human beings, but also across all domains of content"* (Cook & Groom, 2004, p. 37).

"Know the method your subject is using to perform the experimental task" and "never average over methods" (Newell, 1973, p. 293).

Consistent with the various factors and theorizing that leads us to expect participant treatment heterogeneity, there is a growing number of published reports that document heterogeneity for various psychological phenomena:

- Perceptual categorization (Yang & Lewandowsky, 2004)
- Knowledge structure categorization task (consumer context, Blanchard, Aloise, & DeSarbo, 2012)
- Category learning (Lee & Webb, 2005; Navarro, Griffiths, Steyvers, & Lee, 2006)
- Skill acquisition learning (Newell, & Rosenbloom, 1981)
- Memory retention curves (Estes, 1956; Estes & Maddox, 2005; Myung, Kim, & Pitt, 2000)
- Recognition memory (Dennis, Lee, & Kinnell, 2008; Rouder, Lu, Morey, Sun, & Speckman, 2008)
- Verbal short-term memory (Logie et al., 1996)
- Arithmetic performance (Siegler, 1987, 2007)
- Spatial cognitive mapping (storage and processing of geographical information, Kitchin & Fotheringham, 1997)
- Decision making under risk (i.e., utility functions, Schunk & Betsch, 2005)
- Omission bias (Baron & Ritov, 2004; Baron, 2010)
- Social norms on energy conservation behavior (Schultz, Nolan, Cialdini, Goldstein, & Griskevicius, 2007)
- Authority and consensus influence strategy on persuasion (Kaptein & Eckles, 2012)
- Moral judgments (i.e., punishment judgments, Baron, & Ritov, 2009a)
- Big-Five factor model of personality (Borkenau & Ostendorf, 1998)
- Children's emotional reactions to parent interactions (Molenaar, Sinclair, Rovine, Ram, & Corneal, 2009)
- Social distress on willingness to provide support (Whitsett & Shoda, 2014).

In summary, there are many *a priori* reasons to expect heterogeneity in psychological phenomena across individuals, meta-scientific theorizing by our forebears warning us about possible heterogeneity, and a growing number of published articles reporting such heterogeneity.

Averaging across such heterogeneity leads to tiny effect sizes that are prohibitively difficult to detect given typical sample sizes researchers have access to. This is the case for two kinds of heterogeneity: (1) Heterogeneity whereby a proportion of subjects show the effect with the remaining proportion of subjects *not* showing the effect and (2) heterogeneity whereby a

proportion of subjects show the effect, another proportion *not* showing the effect, and the remaining proportion of subject showing an *opposite* effect (tripartite heterogeneity). Averaging across individuals for both of these kinds of heterogeneity – in particular the latter kind – will result in small effect sizes that are prohibitively difficult to detect given typical sample sizes researchers have access to.

For example, for between-subjects designs, a small effect size of d=.20 requires an N=1,302 to reliably detect with 95% power.[1] This becomes even more dramatically prohibitive for interactions. For example, a 2x2 attenuated interaction of equivalent effect size (d=.2) requires an N=2,604 to reliably detect with 95% power (calculations executed via simulations provided by Simonsohn, 2014).[2] Though tiny effect sizes can be important in certain applied contexts (e.g., aspirin heart attack study, marketing studies where small increases translate into millions of dollars, etc.), documenting such tiny average effects is inconsistent with the goals of basic psychological research, which aims to understand at a fundamental level individuals' mental processes and behavior. With these considerations in mind, it is clear that we need a new (modal research) way forward.

New meta-theoretic framework

The new proposed framework provides a general framework for examining psychological phenomena at the level of the individual. The framework offers substantial potential in overcoming psychology's replicability problem by yielding much more nuanced information that significantly increases the probability of observing consistent results across independent samples, but also provides a constructive approach in building a cumulative knowledge base by more clearly indicating avenues for follow-up incremental work (examples below will make this clearer).

From our new meta-scientific perspective, research questions are examined using designs that allow one to probe psychological phenomena separately for each individual in one's sample. Consequently, the framework allows researchers to empirically determine – for a particular psychological phenomenon – the proportion of individuals who:

1. Show a psychological effect (E%)
2. Do not show a psychological effect (N%)
3. Show an opposite psychological effect (O%)

Researchers would then attempt to replicate these proportions in an independent sample drawn from the same population. Once these proportions are determined, an important goal then becomes to identify inter-individual factors (e.g., individual difference variables) that successfully differentiate between (1) the individuals who show an effect versus, (2) those who do not show an effect versus, and (3) those who show an opposite effect. "Showing an effect" refers to any within-person effect from simple one-way (two-condition) main effects, to more

[1]This is a reasonable effect size given meta-meta-analyses yield average effect sizes in psychology of about d=.40 (Lipsey & Wilson, 1993; Richard, Bond, & Stokes-Zoota, 2003), which would likely shrink to about d=.20 after adjusting for publication bias and questionable research practices (John et al., 2012; Simonsohn, Nelson, & Simmons, 2014)

[2]Of course, these numbers are not as prohibitive for within-subjects designs, however, heterogeneity issues are equally problematic for within-subjects designs given that analyses nonetheless average across subjects.

complicated one-way (three-condition) contrasts, to even more complicated two-way interaction effects.

For example, let's consider the previously mentioned scarcity effect from the perspective of our proposed framework. By using an intra-individual design where participants are exposed to a series of consumer products that vary in their perceived scarcity (e.g., using a highly-repeated within-person (HRWP) design, Whitsett & Shoda, 2014), one could empirically determine the (1) proportion of participants who exhibit a scarcity effect (e.g., stronger purchase intentions for scarce compared to control products) and (2) the proportion of participants who do not exhibit a scarcity effect. Then, researchers would seek to identify individual difference variables that can successfully discriminate between individuals who exhibited the scarcity effect versus those who did not (e.g., only individuals who self-report a certain minimum level of need for closure, as suggested by Jung & Kellaris, 2004). The related consensus persuasive strategy effect (Cialdini, 1993) can also be applied to our framework to exemplify the second (tripartite) kind of heterogeneity (see Figure 1, right panel). By exposing participants to a series of books that include different consensus persuasive message (e.g., "Over a million copies sold") versus control messages (e.g., "This book is a random selection from our catalogue"), one can empirically determine the (1) proportion of participants who exhibit a consensus effect (e.g., stronger purchase intentions for best sellers compared to control books; "Subject #1" in Figure 1), (2) the proportion of participants who do not exhibit a consensus effect (e.g., "Subject #3 in Figure 1), and (3) the proportion of participants who exhibit an "anti-consensus" effect (e.g., weaker purchase intentions for best sellers compared to control books; e.g., "Subject #6" in Figure 1). The latter "anti-consensus" cluster of individuals could reflect individuals with esoteric tastes in books who use consensus information as a heuristic suggesting those books are unlikely interesting if they are liked by the mainstream.

In the following section, we more concretely demonstrate the new framework by applying it to two different psychological phenomena (i.e., implicit race bias, affect misattribution). The empirical demonstrations involve real data and showcase intra-individual designs (e.g., HRWP) and analytical tools (e.g., mixed-effects models, latent-class multinomial processing tree models) that can be used to investigate psychological phenomena from this new meta-theoretic perspective.

Empirical demonstrations
(Individual-level analyses yield more consistent & informative results across samples than aggregate-level analyses)

- Demonstration #1: Weapon Identification Task (WIT; Correll et al., 2002)
    - **Agg.-level:** d = .06, p>.46 in sample #1, d = .11, p>.17 in sample #2
    - **Ind.-level:** ~57% Ps show anti-Black bias and this is consistent across samples (see Table 1)

Table 1: Individual-level results across Sample #1 and #2

Sample	**Individual-level Effects**

Sample #1 (N=148)	**#E% (Anti-Black Slopes): 53% (78)** **#N% (Non-signif Slopes): 41% (60)** #O% (Anti-White Slopes): 7% (10)
Sample #2 (N=149)	**#E% (Anti-Black Slopes): 62% (92)** **#N% (Non-signif Slopes): 32% (48)** #O% (Anti-White Slopes): 6% (9)

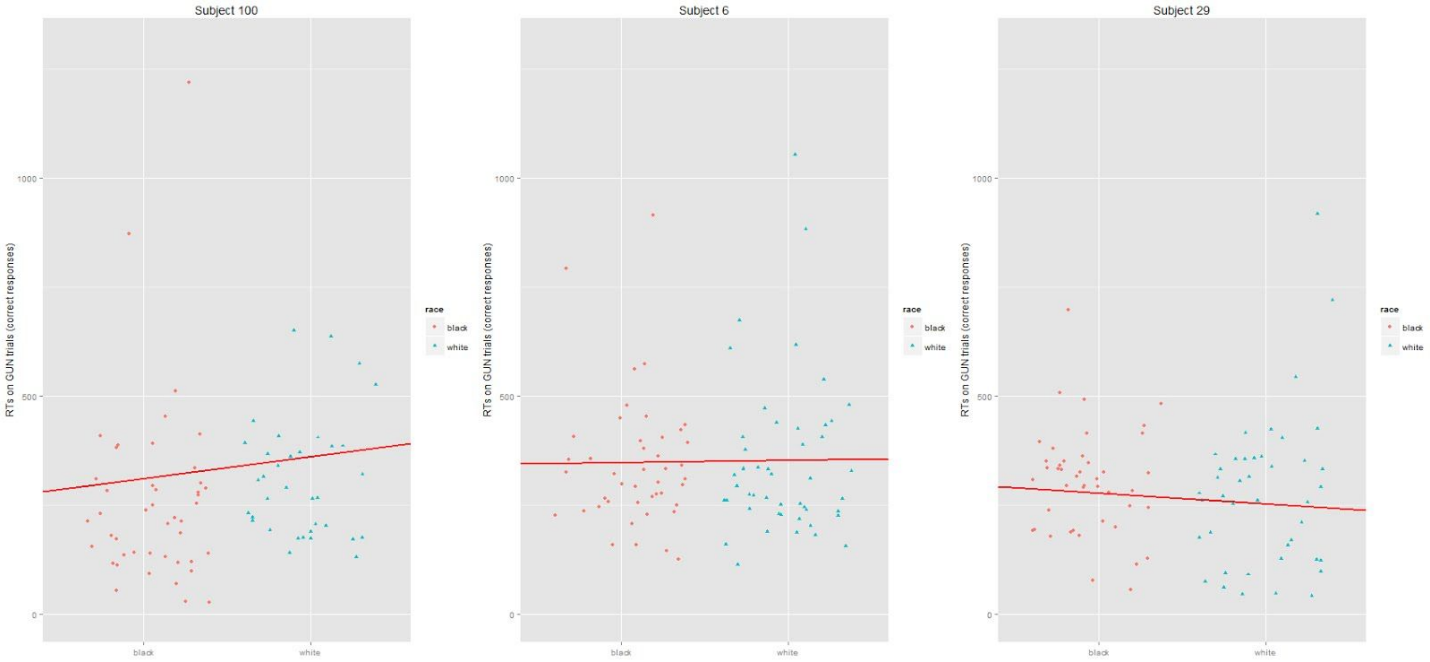

Figure 2. Individual-level results showing an individual exhibiting anti-Black bias (Subject 100, left), not showing bias (Subject 6, middle), and showing a anti-White bias (Subject 29, right).

- Demonstration #2: Affect Misattribution Procedure (AMP; Payne et al., 2005)
  - **Agg.-level:** MPT model not supported in sample #1, model supported in sample #2
  - **Ind.-level:** Latent-class approach reveals more informative and consistent story across samples (see Table 2)

Table 2: Individual-level results across Sample #1 and #2

Sample	**Individual-level Effects**
Sample #1 (N=71)	**61% of Ps show Misattribution & Affective priming** **39% of Ps do not show Misattribution nor Affective priming**
Sample #2 (N=75)	**78% of Ps show Misattribution & Affective priming** **22% of Ps do not show Misattribution nor Affective priming (actually show reverse Affective priming)**

Furthermore, only 31% and 40% of participants showed a statistically significant priming effect in samples #1 and #2, respectively, whereby pictographs were evaluated more positively when preceded by positive compared to negative prime photos. Such evidence suggests that "affect misattribution" in the AMP is unlikely to be a general mechanism underlying the AMP given that the majority (i.e., ~65%) of individuals do not even show the necessary (but insufficient) affective priming effect.

Applied Examples

Applying framework to other phenomena in other areas of psychology

Area	**Phenomenon**	**Concrete Example**
Social	Personalized prejudice reduction Disgust on moral judgments	
Industrial/Organizational	Incentives on motivation/productivity (e.g., positive/negative bonuses)	
Personality	Reducing SDR (compare techniques, e.g. BTS vs. RRT?)	
Positive psychology	Gratitude diaries	
Educational	Learning styles Note-taking method	
Clinical	Expressive writing	
Health psychology	Binge drinking Food cravings (compare interventions)	

Concerns/Misconceptions

In this section, we clarify some anticipated misconceptions and concerns regarding our proposed approach.

- *Misconception #1:* But we do account for individual idiosyncracies via individual difference moderators
    - Simply not true; see Figure 1 and provide another concrete example

- *Misconception #2*: This is impossible for most research questions/phenomena in psychology
    - Though it's certainly true that the proposed intra-individual approach *is* impossible for certain psychological phenomena involving permanent psychological changes (e.g., certain kinds of learning), we contend such

phenomena are the exception rather than the rule. Hence, our approach still has much potential in a general sense.
    - Ultimately we won't know until we try; science is one of the most challenging endeavors that requires extreme levels of creativity and innovation
        - Many phenomena that seemed/were impossible to measure/study, eventually turned out to be measurable/studyable many years later in the wake of innovations)

This notion is very well captured in the following passage by the great Paul Meehl (1990):

This misconception is that, if a theoretical conjecture is "scientifically meaningful"…, then it must be possible to test it at the present time. Even a slight familiarity with the history of astronomy, physics, chemistry, medicine, and genetics shows that such a meta-theoretical notion is plainly false. These other sciences are replete with examples of perfectly good "empirical" questions, askable by sophisticated scientists at a given time, that could not be answered given either deficiencies in the required auxiliary theories or the lack of an adequate instrumentation, whether for control of variables on the one side or, more commonly, measurement of variables on the other (p. 238-239).

Implications/Future Directions

- Calls for more sophisticated methods (e.g., HRWP, HLCMPT) for studying complex (social) psychological phenomena including:
    - masking techniques to minimize demand effects in within-person designs
    - more advanced, systematic, and standardized forms of funnel debriefing
    - larger samples sizes at both the subject *AND* "trial" level

- Approach can and eventually should be extended to "occasions" across time (i.e., in addition to probing ergodicity in effects across individuals (what our main focus is on), also likely need to probe ergodicity in effects across time within-person for many phenomena)
- Also – and almost like a 3rd step to the 2-step approach described above – once heterogeneity in effects are empirically determined, ultimately we need to do further work to understand the *mechanisms* underlying within-person effects (because of equifinality; i.e., two different individuals can show the same effect due to two distinct mechanisms!)

Conclusion

- To overcome its replicability crisis, experimental psychologist need to investigate -- and test hypotheses about -- psychological phenomena at the level at which these phenomena exist: at the level of the individual. Then, and only then, will experimental psychology have any hope of developing a cumulative, valid understanding of psychological phenomena.

References (incomplete)

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A, Fiedler, K, …, Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27,* 108-119.

Bakker, M., van Dijk, A, & Wicherts, J. M., (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*, 543-554.

Chambers, C. D. (2013). Registered Reports: A new publishing initiative at Cortex. Cortex, 49, 609–610.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160.

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives in Psychological Science, 7,* 645–654.

Koole, S. L., & Lakens, D. (2012). Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science, 7*, 608-614.

LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology, 15*, 371-379.

LeBel, E. P., Borsboom, D., Giner-Sorolla, R., Hasselman, F., Peters, K. R., Ratliff, K. A., & Smith, C. T. (in press). PsychDisclosure.org: Grassroot support for reforming reporting standards in psychology. *Perspectives on Psychological Science.*

Makel, M. C., Plucker, J. A., & Hagerty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science, 7*, 537-542.

Neuliep, J. W., & Crandall, R. (1990). Editorial bias against replication research. *Journal of Social Behavior and Personality, 5*, 85–90.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7,* 615-631.

Open Science Collaboration, T. (2012). An open, large–scale, collaborative effort to estimate the reproducibility of psychological science. Perspectives on Psychological Science, 7, 657–660.

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7*, 528-530.

Popper, K. R. (1959). *The logic of scientific discovery.* Oxford, UK: Basic Books.

Rosenthal, R. (1997). Some issues in the replication of social science research. *Labour Economics*, *4*(2), 121-123.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16,* 225-237.

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods, 17*, 551-566.

Simonsohn, U. (2014, March 12). No-way interactions [Blog post]. Retrieved from http://datacolada.org/2014/03/12/17-no-way-interactions-2/

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology, 100,* 426-432.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J. & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*, 627-633

Wicherts, J. M. (2013). Science revolves around the data. *Journal of Open Psychology Data, 1(1)*,1-4. DOI: http://dx.doi.org/10.5334/jopd.e1

Wolfe, J. M. (2013). Registered reports and replications in Attention, Perception, & Psychophysics. Attention, Perception, & Psychophysics, 75, 781–783.