Grounding Applied Social Psychology in Translational Research

Klaus Fiedler (Universität Heidelberg)

Klaus Fiedler

(University of Heidelberg)

Running head: Grounding applied social psychology

Author Note: The work underlying the present article was supported by a grant provided by the Deutsche Forschungsgemeinschaft to the first author (FI 294/26-1). Email correspondence may be addressed to kf@psychologie.uni-heidelberg.de

Abstract

An often-noted truism says that fundamental and applied research can complement and fertilize each other. A less common insight is that translational research conducted in the lab under experimentally controlled conditions can inform as practical interventions as applied field research conducted under naturalistic conditions. To illustrate this challenging and somewhat provocative point, in the present paper I present several examples of important practical insights and successful interventions that were inspired by abstract ideas derived from sampling-theoretical approaches to rationality research. Consistent with Kurt Lewin's (1943) notion that "there is nothing as practical as a good theory" [p. 118], translational research findings obtained under controlled lab conditions often result in more well-understood and more clearly specified predictions and interventions than applied research findings obtained in the wild.

Introduction: From "Applied" to "Translational" Research

Kurt Lewin's wisdom that there is nothing as practical as a good theory is frequently cited and it seems to be endorsed and embraced by almost everybody who is citing this statement. However, what does the statement mean? Did Lewin intend to say that when basic research is inspired by a good theory, it can be applied immediately to the real world, or that good theories allow us to equate theoretical and practical research? For instance, can the finding that distributed learning is superior to massed learning (Kornell & Bjork, 2008), which derives from an approved, incontestable theoretical principle, be equated with practical improvements in academic settings and everyday learning practices? Or, does the growing body of recent evidence on the "wisdom of crowds" (Surowiecki, 2004), which is logically rooted in Bernoulli's (1713) law of large numbers imply that, practically, all groups are more rational than all individuals, and that the accuracy of all group decisions increases with group size (Sorkin, Hays & West, 2001), or that test validity always increases with the number of test items?

The answer to all these questions is clearly: No. A good theory cannot be simply equated with a successful practical intervention. There is still a long way from having a good theory to achieving practical success in applied settings. Lewin certainly did not want to downplay the amount of work and will-power required to translate a fully convincing idea into practical benefits. And, it might therefore be more adequate to refer to translational science rather than to applied science. To quote from Wikipedia (https://en.wikipedia.org/wiki/Translational_research): "translational research applies findings from basic science to enhance human health and well-being". It has to be admitted, though, that there are countless examples of good theories that were never translated into improved practices. The advantage of distributed learning was never implemented into university curricula. The wisdom of crowds does not prevent democratic decisions from being often biased and irrational. Or, to provide another memory example, more

than a hundred years after Galton's (1907) insightful work on regression effects, the "regression trap" continues to fool laypeople and experts in all areas of applied science (Campbell & Kenny, 1999; Fiedler & Krueger, 2012; Fiedler & Unkelbach, 2014)). That is, the failure to acknowledge that in a noisy world, plotting a second measure as a function of a primary measure must always produce a regression line with a slope less than one,[1] is at the heart of countless misunderstandings and irrational decisions and actions.

A more moderate and realistic interpretation of Kurt Lewin is still encouraging: Successful practical interventions represent an ambitious and rarely accomplished goal, the attainment of which is however greatly facilitated when translational research is anchored in a clearly spelled-out theory. Some impressive examples of successful translational research substantiate this assumption. Fundamental research on face recognition in multiple-choice settings have led to marked procedural changes and improvements in eyewitness-identification procedures (Wells et al., 2000; Wixted & Wells, 2017). For example, theory-driven memory research has shown that sequential lineups are more likely than simultaneous lineups to produce false positives, that is, false identifications and convictions of innocent suspects. As a consequence, sequential procedures have been implemented in many places. More generally, translational research has produced profound knowledge of how to render lineups pristine diagnostic procedures (Wixted & Wells, 2017): only one suspect per lineup; suspect never more salient than fillers; caution that lineup might not include offender; double-blind testing; online confidence statement at the time of testing. Note that all these features of a pristine lineup are imported from basic-research methodology, following the notion of "line-ups as experiments" (Wells & Luus, 1990).

[1] For example, replication results must always be less pronounced than original results, for purely logical reasons (Fiedler & Prager, 2018).

Other telling examples of uncontested practical success due to theory-anchored psychological science include, for instance, Johnson and Goldstein's (2003) memorable demonstration that willingness to become an organ donator depends to an amazing degree on legal default setting. In countries like Austria and France, in which people are by default registered as organ donators, unless they actively declare to decline, almost all citizens do not change their default status and are therefore ready to donate their organs after they die. In contrast, willingness to be registered as organ donator is conspicuously low in countries like Germany or the Netherlands, where the default setting is not to donate.

Another nice example would be Ritov's (1996) translational work on anchoring effects in competitive markets (such as Ebay). This research has shown that by far the best predictor of the final selling price in an auction is the initial offer, indicating that too much modesty at the beginning is hardly a prudent strategy. It is interesting to note that even exceptional evidence that exaggerating starting claims may sometimes discourage competitors and thereby reduce the final price can be well understood in a theory-driven translational research framework.

Note, however, that a good theory originating in fundamental science not only increases the chances of strong practical interventions. One could even argue that controlled lab experiments and refined models developed in basic research can be practically more useful than any attempt to conduct "practical research" in the real world. Legal, ethical, and pragmatic constraints do not allow us to induce real phobia and depression, to expose participants to dangerous risks, to let people gamble with non-trivial sums of money, to induce strong emotions with serious side-effects, to apply different teaching methods to different school classes, or even to access confidential personal data from therapy, negotiation, or decision protocols. Moreover, because internal validity always sets an upper limit for external validity (Campbell, 1957), scientific hypothesis testing under controlled lab conditions must be logically and pragmatically antecedent

to testing the generality and the limitations of hypotheses under natural conditions. As Bjork (1994) has shown impressively, the results of research that attempt to mimic reality in the lab may lead to less transferrable evidence than the deliberate manipulations of theoretically motivated interventions under controllable research conditions.

Using Sampling-Theory Approaches to Inform Translational Social Psychology

In the remainder of this article, I try to demonstrate and illustrate the emergence of translational research as a natural side effect of basic research in a theory domain that has been shown to produce "good theories" on the assembly line, namely sampling approach to judgment and decision making. Rather than providing further post-hoc examples of influential applied research that can be traced back to basic research, the strategy here is the other way around, namely, to start from cogent theoretical ideas that promise to have strong practical implications. It goes without saying that such practical implications are still waiting to be translated, cross-validated, and exploited in applied domains. However, the applied-psychological potential should be vividly apparent in any case.

Sampling approaches are well suited for at least two reasons. First, statistical sampling theory imposes a number of strong constraints on psychological predictions, which are derived on solid logical or mathematical ground. So, in the realm of sampling theories (Fiedler & Juslin, 2006; Fiedler & Kutzner, 2015) there can be no doubt about what a theory really implies. Second, the very samples of observations supposed to mediate rational or irrational judgments and decisions can be assessed in a typical sampling study, independently of the behavior to be explained. So they provide excellent opportunities to validate the supposed mediation process.

Baserate Neglect and the Asymmetry of Conditional Samples

Let us start with an illustration of this crucial point, using an example related to the so-called base-rate neglect (Bar-Hillel, 1984; Tversky & Kahneman, 1974). An uncontestable

property of the empirical world is that sample estimates of conditional probabilities can be highly asymmetric. The conditional probability *p*(blood alcohol|accident) that if a car driver is involved in an accident, he or she has an elevated blood alcohol level is presumably much higher than the reverse conditional probability *p*(accident|blood alcohol) of an accident given elevated blood alcohol. Assuming that the base-rate of car accidents is (fortunately) much lower than the base-rate of elevated blood alcohol – that is, *p*(accident) < *p*(blood alcohol) – the former conditional must be much lower than the latter conditional: *p*(accident|blood alcohol) < *p*(blood alcohol|accident). The general mathematical (Bayesian) rule says that the ratio of conditionals is equal to the ratio of base-rates, *p*(Y|X) / *p*(X|Y) = *p*(X) / *p*(Y). Thus, the conditional probability of a less likely event given a more likely event must be lower than the reverse conditional probability of a more likely event given a less likely event.

However, laypeople like experts often exhibit a base-rate neglect, that is, their judgments fail to take the asymmetry of base-rates into account, especially when the relevant conditional is less apparent or is not available at all (Gavansky & Hui, 1992). For instance, when estimating the danger of blood alcohol, which calls for an estimate of *p*(accident|blood alcohol), the relevant statistics are not available, because one cannot assess the level of blood alcohol in a representative sample huge enough to yield a reasonable number of such rare events as car accidents. In such a situation, people commonly rely on the reverse conditional, *p*(blood alcohol|accident), which is easily available because virtually all drivers involved in an accident undergo an alcohol test. It should be obvious, though, that *p*(blood alcohol|accident) must grossly overestimate the danger of blood alcohol. Assuming that elevated blood alcohol is, say, ten times more likely than accidents, the available sample statistic exaggerates the danger by the same factor 10!

Self-evident and banal as this example may appear, it is at the heart of some of the strongest biases in risk assessment. Because the base-rate *p*(HIV) of people who have contracted a HIV virus is much lower than the reverse base-rate *p*(+tested) of people who are tested positively on an HIV test, the probability *p*(HIV|+ tested) that a positively tested person has the virus is by magnitudes smaller than the reverse probability (hit rate) that a carrier of the HIV virus is tested positively. Whereas the latter conditional is virtually perfect, *p*(+ tested|HIV) ~ 1.00, the former conditional is as low as *p*(HIV|+ tested) ~ 0.15, as nicely explained by Swets, Dawes, and Monahan (2000).

Presumably, the diagnosticians' strong tendency to avoid false negatives (i.e., to overlook cases of real HIV), maybe for liability reasons, has led to the implementation of diagnostic tests that are overly sensitive, producing clearly more positive test results than there are HIV cases in the population. As a consequence of such base-rate neglect (Bar-Hillel, 1984), a huge number of false positives leads practitioners and lay people to drastically overestimate the certainty with which a positive test indicates the HIV virus. To repeat, if someone is tested positively in an unselective screening test, the true probability that the person actually has the virus is as low as 15%![2]

However, crucially, within a sampling-theoretical framework, it is possible to elucidate the process leading to such dangerous judgment biases. Thus, Fiedler, Brinkmann, Betsch and Wild (2000) provided participants with an index-card file of patients, with the diagnosis (breast cancer or no breast cancer) on one side and a mammography test result (positive vs. negative) on the

[2] When I included this evidence, which is based on authentic data analyzed by Swets et al. (2000), the editor wanted to intervene because of friend of his, an experienced radiologist, told him these statistics cannot be correct. However, the figures are valid. The problem is only that even experts like the editor's radiologist friend fall prey to base-rate fallacies.

other side of each patient's index card. The distribution of all 2 × 2 combinations of diagnoses and test results corresponded roughly to the actual base-rates and conditionals in the population. However, because the two base-rates, *p*(breast cancer) < *p*(positive mammogram), are highly unequal (as in the HIV example above), the final judgments of *p*(breast cancer|positive mammogram) were radically different in two sampling conditions. When in one condition the index-card file was organized by test results, participants understood that they only had to draw a sample from the slot with positive test results, and the observed proportions of breast cancer cases indicated on the back side of the sampled cards provided a rather precise estimate of the conditionals in question. In contrast, when in another condition the index-card file was organized by diagnosis, offering a very large number of no-breast cancer cases in one slot but only very few cases of breast cancer in the other slot (corresponding to the low base-rate), participants would typically draw all breast cancer cases along with roughly the same number of no-breast-cancer cases. Thus, when the relevant conditional was not easily available, the reverse conditionals misled participants to provide drastic overestimates of *p*(breast cancer|+ mammogram) based on samples that contained breast cancer at a rate as inflated as 50% (i.e., inflated by 5000%)!

Recall that this approach to translational research on risk assessment does not rely on weak speculations but on uncontestable Bayesian probability theorems. It does not run the danger of producing wrong predictions or non-replicable results. Not surprisingly, these results have been replicated again and again (Fiedler, 2008; Fiedler et al., 2000; Fiedler, Hütter, Schott & Kutzner, 2019) under varying boundary conditions. And most importantly for Lewin's notion of a "good theory", it can be shown that the process underlying the sign and size of faulty risk assessments (Bodemer & Gaissmmaier, 2012) can be convincingly explained by the objectively available sampling data. The strength of individual judges' biases is clearly predictable from the strength of the biases inherent in the individual judges' conditional samples drawn from the index-card file.

Output-Bound Sampling

The research documented in Fiedler (2008) highlights the malicious role of output-bound sampling. Participants were asked to take the role of a leading manager whose task is to purchase two kinds of equipment, computers and telecommunication devices, offered by three providers. To make optimal purchasing decisions, they could sample information from an available data bank about prior customers' positive (+) versus negative (–) experience with all three providers, with regard to both product types. On every trial, they could specify as many aspects as they wanted, that is, they could either ask for a new random draw from all database entries pertaining to computers from Provider 1, or simply computers from any provider, or negative experienced with telecom devices, or any outcome and product from Provider 2, or they could leave all aspects open and simply ask from the next random draw from the entire database. In fact, in the entire database, the true success rates were exactly the same across all providers and product domains. Only the baserates varied; the frequency ratio of positive to negative experiences was constantly 2:1 for all three providers and for both product domains. However, computer entries were twice as frequent as telecom entries, Provider 1 was twice as frequent as Provider 2, who was twice as frequent as Provider 3.

Only by never (not even on a subset of trials) specifying any provider, domain, or experience condition, but by confining oneself to an unconditional sample from the whole database could participants obtain an unbiased sample (as in the upper chart of Figure 1). Unsurprisingly, though, such a strategy was virtually never chosen by the participants. They rather focused on those conditions that were relevant to testing specific hypotheses. For instance, when asked to find out whether Provider 3 should be preferred to Providers 1 and 2, they tended to gather relatively more observations about the focal Provider 3, thus producing a sample as in the middle chart of Figure 1. Or, when the task called for diagnostic inferences about the sources

of deficits or negative experiences, they would conditionalize negative outcomes and leave providers and domains open, receiving a sample as in the bottom chart.

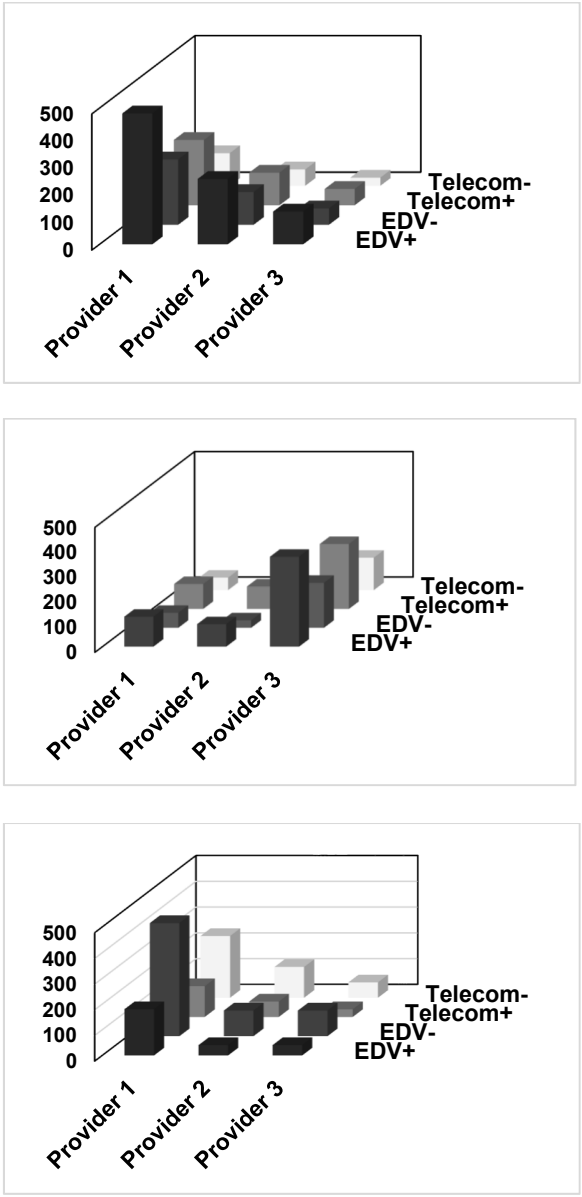

Figure 1: Frequency distribution of positive and negative experience with two product types (computers and telecom) from three different providers. The upper chart shows an unbiased (unconditional) sample; conditional samples from the same universe look quite differently when comparing Provider 1 with Providers 2 and 3 (middle chart) or when trying to detect the origins of negative experiences, or deficits (lower chart).

The final judgments and purchasing decisions were reflective of the very sampling biases. Samples of the first two types would reveal that most experiences were positive, but such positive outcomes were more strongly associated with Provider 1 in the upper chart but relatively more associated with Provider 3 in the middle chart. In contrast, when diagnosing deficits, sampling concentrated on negative outcomes, and the more frequently observed providers became the most frequent origins of the over-sampled negative outcomes. In any case, these distinct biases in the resulting preference judgments were predictable from the relative strength of the sampling biases experienced by participants who pursued different hypotheses.

Thus, although judgments were highly sensitive to the sampled stimulus distributions in a three-dimensional space (providers × product domains × valence), they arrived at completely different evaluations, depending on the sampling perspective induced by the task instructions. An analysis of the underlying sampling processes revealed – thanks to the "good theory" that guided the entire investigation – one conspicuous source of highly irrational judgments biases: output-bound sampling. Output-bound sampling occurs when information search is selectively guided by the variable to be estimated from the sample. Thus, to diagnose the causes of product deficits, the sampling process focuses on mostly negative outcome, and based on the resulting (predominantly negative) sample, the most prevalent provider is found to be very negative. But this negativity reflects the judge's own sampling bias. Although the most likely outcome in the database is clearly positive, the deliberate focus on the analysis of deficits produces highly biased samples that strongly over-represent negative outcomes, and this self-determined sampling bias misleads judges to provide mostly negative judgments of the provider with the highest base-rate.

Practical importance of output-bound sampling biases in applied contexts. Output-bound sampling is prototypical of translational research. An abstract concept from fundamental

research turns out to be of enormous practical value in many applied domains. The likelihood p(HIV | positive HIV test) that a person who is tested positively actually has the HIV virus is strongly overestimated because the samples used for the estimate contain roughly the same 50% of HIV cases and non-HIV cases, even though the HIV base-rate in the universe is much less than 1% (Fiedler, Brinkmann, Betsch & Wild, 2000; Fiedler et al., 2019). To be sure, the estimates accurately reflect the proportion of HIV cases in the sample, but the sample on which the estimate is based grossly over-represents the variable to be estimated.

Some manifestations of output-bound sampling appear blatant and almost unbelievable. For instance, the continued high reputation of polygraph lie detection is due to the systematic exclusion from study samples of those cases that could have invalidated the polygraph test. Only cases in which a positive polygraph test is validated by a defendant's (genuine or strategic) confession are typically retained in the sample, because a confession is considered to be essential for a criterion of the "ground truth" (Fiedler, Schmid & Stahl, 2002). Innocent cases that do not confess are thus excluded by a sampling strategy that comes close to self-deception.

Another example from legal psychology concerns the composition of lineups for eyewitness identification. A severe case of output-bound sampling arises when the selection of a suspect for a lineup is based on a synthesized photo constructed from the witness' own report of the perpetrator's appearance. Given such output-bound sampling of the suspect's appearance, completely tuned to the witness' memory bias, it is no wonder that the same witness then often identifies the (innocent) suspect that resembles the witness own memory biases (see Wixted & Wells, 2017, for a memorable review).

"Metoo" Movement. For a particularly prominent recent example of a media effect, consider the so-called *Metoo movement*, that is, the snowball effect of people revealing that they have also been the victim or rape or sexual assault or harassment. There can be no doubt that this

public sensitization movement serves the function of uncovering intolerable and often criminal behavior and that we have to do all we can to tackle this problem. Nevertheless, the Metoo revelation process is seriously biased as an estimation procedure, because it is built on a severe form of output-bound sampling, mobilizing exactly the subset of sexual victims, whose prevalence cannot be estimated from such a sample. Nevertheless, the absolute volume of the sample of women joining the Metoo movement is interpreted as an indicator of the prevalence of a serious social problem. This is not only irrational and strongly misleading, because the output-bound sampling design does not allow us to make any inferences about relative prevalence. One may also suspect that inflated estimates of sexual transgressions create a descriptive norm (Cialdini, 2012) that serves to downplay the severity of apparently "normal" habits.

Pitfalls of sampling truncation. An intriguing special case of output-bound sampling arises when sample size is not an independent variable but when the stopping rule depends on the information gathered so far. For example, an interview or interrogation is truncated as soon as sufficient information has been accrued; consumers cease sampling information and draw a purchasing decision when the know enough about different products; a democratic decision is made when a committee has formed its preferences; or an empirical study is closed at the moment the results look optimal. Under such conditions of self-truncated information sampling, the law of the large number turns into the opposite, a small-sample advantage (Prager & Fiedler, 2019; Prager, Krueger & Fiedler, 2018). A self-truncated small sample has remained so small exactly because it happened to provide such a clear-cut picture of the existing trend in the universe. In contrast, a larger sample had to grow to such a size because it did not reveal a clear-cut impression from the beginning. Thus, when sample size becomes a dependent variable that reflects the amount of information or diagnosticity of the initial evidence, a characteristic "less-is-more" effect can be exploited (Gigerenzer, Todd & the ABC group, 1999; Hogarth & Karelaia,

2005; Katsikopoulos, 2010). Quick actions and decisions can be made at high confidence levels based on small samples causing little conflict and low information costs.

Output-bound sampling in empirical research. Because trust in the reliability and usability of scientific results is an essential aspect of translational science, it seems appropriate here to include a note on misleading inferences from replication research. The recent debate about the provocatively low replication rates of research findings from psychology, medicine and other disciplines (Camerer et al., 2018; Ioannidis, 2005) that have been reported in several frequently cited articles are to an unknown degree contingent on output-bound sampling of findings that can be expected to be unlikely, unexpected, original, not overlapping with previously published findings and – notably – not based on a good theory. Papers published in such leading journals as Science, Nature, Psychological Science, have to meet several criteria that render them hard to replicate and good candidates for replication failure: They have to be sexy, hard to believe, and most importantly, not merely replicating what is already established in cumulative science. Moreover, many published studies are not designed to maximize external validity (Campbell, 1957). It is not surprising therefore that many of these novel and unexpected findings turn out not to reflect a universal law that generalizes beyond the original conditions that let them appear sexy. In other words, the premise that every published finding constitutes a candidate for replication may not be warranted. It is interesting in this regard that although only 3x% of a sample of findings published in Science and Nature were classified as replicable by Camerer et al. (2018), a vast majority of all the authors correctly identified the small subset of replicable findings. Apparently, then, the very sample is contaminated by output-bound sampling, containing predominantly findings that nobody would expect to be replicable.

Sampling at Different Aggregation Levels

Success and failure of evidence-based decision-making and political intervention depend on the analysis of the available data at an appropriate level of aggregation. Different aggregation levels can convey radically different pictures of the same reality. At the individual level, the rate of illiteracy is almost the same for Black and White people, but at the level of districts or social ecologies, the correlation between illiteracy and Black versus White race can be as high as $r = .92$ (Robinson, 1950). Economic wealth at the national level (in terms of the national gross product) can come along with a very high rate of poverty at the individual level. Or, the same consumer products may be sold at higher prices in cheaper than in more expensive supermarkets (Vogel, Kutzner, Fiedler & Freytag, 2013).

As a consequence, many political debates and economic conflicts arise because different parties rely on different aggregation levels, often depending on which level provides a more favorable picture of one's own position. For instance, a rightist political party may praise the seemingly successful gross product in a country that minimizes taxes for industrial organizations, whereas a leftist party in the same country may complain about high individual-level rates of unemployment and poverty. Or, group-level preferences may diverge from the individual-level preferences in democratic decision making (Fiedler et al. 2015), just as group-level research data may not tell us much about the behavior of individual participants in behavioral research.

The basic insight that different aggregation levels often reflect different causal mechanisms is of utmost importance for all applications of behavioral research. The danger of confusing highly divergent trends that may coexist at different aggregation levels is particularly striking in recent analyses of "big data", which come with the alleged guarantee of high reliability and robustness. Suffice it to illustrate this point here with reference to a couple of recent studies of the relationship between air pollution and unethical behavior. One investigation by Lu, Lee, Gino, and Galinsky (2018), based on air pollution data and rates of six crime types in 9360 American

cities, found that all crime rates and all pollution indices jointly decreased regularly from 1999 to 2009. This led the authors to infer a systematic positive relationship between air pollution and ethical behavior. In contrast, in another recent big-data study using monthly data from British cities, the authors found that air pollution is regularly stronger in winter months whereas the prevalence of different crime types increases strongly in summer time (Heck, Thielmann, Klein, & Hilbig, 2019). This pattern shows a marked negative relationship between pollution and immoral behavior. Not surprisingly, annual data reflect completely different causal influences (e.g., legal and political changes across an entire decade) than monthly data (e.g., seasonal temperature changes). At the same time, this example highlights the insight that the one and only true correlation between air pollution and immoral behavior does not exist. Different correlations seem to reflect different underlying processes at different levels of aggregation, calling for "good theories" that are sorely needed to understand the evidence gained from "big data".

Evading the Regression Trap in Translational Research

Let us finally consider the practical significance of the "regression trap" – another topic of basic and translational research that applied social psychology must keep in mind. Regressiveness is an essential property of the probabilistic world, which is however notoriously ignored or at least neglected. When predicting one variable Y from another variable X, an imperfect correlation $|r_{XY}| < 1$ implies that predicted Y values must be less extreme than predictor values of X. That is, a regression slope of $|b| < 1$ implies that variance in the predicted values will be less than the variance in the predictor. To estimate the expected deviation $y = Y–M_Y$ of Y from the mean M_Y, the corresponding deviation score $x = X– M_X$ of the predictor has to be multiplied by the correlation r_{XY}. If $r_{XY} =0.5$, the extremity or deviation scores of Y will shrink by one half; it $r_{XY} =0.67$, Y deviations shrink to two thirds of the predictor values. Because the principle of regression is well understood, it makes for the kind of "good theory" that Lewin (1943) found to

practical. Indeed, the failure to understand regressive shrinkage renders applied science worthless.

Overconfidence. Let us illustrate this point with respect to one of the most consequential topics in applied decision making science, the so-called overconfidence bias. Let X be subjective confidence that a binary judgment or choice is correct and let Y be the objective accuracy rate. Granting that accuracy is not perfectly correlated with confidence, plotting accuracy as a function of confidence will be subject to regressive shrinkage. Assuming $r_{accuracy,confidence} = .50$, it can be expected that high confidence scores of, say, .40 or .30 above the midpoint will regress to accuracy scores of only .20 and .15 above the midpoint. Because the vast majority of all overconfidence studies plot accuracy as a function of confidence, overconfidence evidence reflects to large extent a regression artifact. As Erev, Wallsten, and Budescu (1994) have shown, the same data set that seems to exhibit overconfidence can be shown to yield underconfidence, when in a reverse analysis confidence is plotted as a function of accuracy. Then very high or high accuracy scores turn out to predict more moderate confidence scores. Ignoring this fundamental insight renders applied research faulty and incompetent. Genuine overconfidence must exceed the normal regression effect that can be expected from the less than perfect confidence-accuracy correlation.

Regression and unrealistic optimism. Because the degree of regression is an inverse function of reliability (r_{XY}), and because reliability increases with amount of information, many practical issues are subject to differential regression. For instance, self-judgments can be predicted to be less regressive than judgments of others, simply because a larger amount of information about the self raises the reliability to a higher level, compared to the lower amount of information about others. Pursuing this hardly contestable idea, Moore and Healy (2008) have greatly developed our understanding of unrealistic optimism, or the often-cited phenomenon that

most people seem to be better than average. However, a more refined analysis informed by differential regression suggests how important it is to distinguish between three different variants of self-related overconfidence: (1) performance *overestimation* relative to objective performance; (2) *overplacement* of oneself rather than others, and (3) *overprecision* of subjective judgments relative to an actually much broader confidence interval.

In general, all judgments, whether they refer to the self or to others, overestimate the (actually low) true performance on difficult tasks and underestimate the (actually high) true performance on easy tasks, in line with the ubiquitous regression effect. However, this so-called hard-easy effect is less pronounced for self-referent than for other-referent judgments. Thus, the typical underestimation bias on easy tasks is stronger for others than for the self. Yet, on difficult tasks, overestimation is more pronounced in other-referent than in self-referent judgments. As a consequence, comparisons of self and others exhibit overplacement (i.e., relatively more positive self than other judgments) on easy tasks but underplacement (i.e., relatively less positive self than other judgments) on actually difficult tasks. Finally, overprecision is more stable than the other two types of overconfidence. Regardless of whether hard or easy tasks produce (optimistic) overestimation versus (pessimistic) underestimation errors, or (self-devaluating) underplacement versus (self-serving) overplacement effects, subjective confidence intervals are generally too narrow. As reported by Moore and Healy (2008), 90% confidence intervals often contain the correct answer at a rate less than 50% (see also Juslin, Winman & Hansson, 2007). Apparently, then, subjective estimates of upper and lower boundaries of a confidence interval are also subject to "normal" regression. Subjective estimates are less extreme than the actual interval boundaries.

The ubiquitous regression trap. A plethora of evidence on biases in laypeople and experts falling into the regression trap highlight Campbell and Kenny's (1999) conclusion that regression is "as inevitable as death and taxes" [p. ix]. Its practical significance can be hardly overstated. For

a very prominent example, the effectiveness of psychotherapy is often over-estimated when interventions start in a crisis, which presumably exaggerates the patient's true pathological state (Campbell, 1996). Because the reliability with which a crisis can be measured is less than perfect, the apparent crisis may to some extent reflect an outlier and the patient's true state can be expected to be less critical. As a consequence, mere regression toward a less exaggerated measure of well-being can underlie an apparent therapy success.

For another example, the stock market is essentially regressive. Granting that stock prices are subject to stationary stochastic variation over time, a wise strategy is to invest anti-cyclically, that is to buy stocks that are in a crisis and to sell stocks that are performing well. By analogy, empirical findings that yielded the strongest effect sizes in the original study produce relatively smaller effect sizes in a replication study, as demonstrated by Fiedler and Prager (2018) with respect to the 100 replications of the Open Science Collaboration, 2015.

Stelzl (1982) discusses the question of why older teachers tend to give up the idealist educational theories of which they were fully convinced when they were young. As teachers get older, they replace their belief in the superiority of reward by the conviction that punishment is more effective than reward. This pessimistic shift may reflect a continuous influence of the regression trap: Assuming that student conduct underlies stochastic variation, positive conduct that is rewarded is typically followed by less positive conduct, whereas negative conduct that is punished is typically followed by more positive conduct. The resulting impression that punishment is more effective than reward may simply reflect the teacher's blindness for regression: Even in the absence of reward or punishment, positive states would have been followed by more negative conduct and vice versa.

Concluding Remarks

In a volume devoted to applied social psychology, the value of solid theorizing cannot be emphasized too much. While Kurt Lewin's (1943) famous parole concerning the practical value of a good theory is often cited and applauded to, it is rarely explained explicitly why this is the case. Providing a twofold answer to this intriguing question was the purpose of the present chapter. On one hand, it was argued and illustrated that translational science – anchored in theory-driven fundamental research in the lab – often entails the potential to be translatable to diverse domains of real social life. Prominent examples include the enormous contributions made by memory researchers to improving eyewitness identification procedures (Wixted & Wells, 2017), the translation of basic insights on effective learning to academic learning contexts (Metcalfe & Kornell, 2007), the key role of base-rate neglect for risk assessment and risk communication (Fiedler, 2010), or the translation of anchoring effects to the applied domain of auctions and negotiations (Ritov, 1996). While the focus in this chapter was on the practical value of sampling theories in particular (Fiedler & Kutzner, 2015), a similar point could be made for practical implications and applications of several theoretical approaches, such as dissonance theory (Cooper, 2012), construal-level theory (Trope & Liberman, 2010), regulatory-focus theory (Higgins, 2012), or theories of affect regulation (Forgas & Ciarrochi, 2002).

On the other hand, however, Lewin's argument must not be reduced to generously granting theory a minor role in a field of applied psychology, which is otherwise dominated by "real" research conducted in the field, with reference to natural situations, political conflicts, social movements, migrants, patients, and victims of aggression and crime. Rather, a more offensive interpretation of Lewin is that such seemingly ultimate applied work under naturalistic conditions can be error-prone, premature, and irresponsible if applied researchers fail to do their homework and take the warnings and insights from fundamental and translational research into account. Thus, an attempt was made to explain that the failure to beware of such theoretical and

methodological issues as the output-bound sampling, the regression trap, and the pitfalls of different aggregation level can undermine the value of even the most laborious and well-motivated applied work.

Finally, and reminiscent of the intriguing lesson conveyed by Bjork's (1994), strictly controlled lab research under control conditions may be more transferrable to the real world than any attempt to mimic "real life" in a controlled study. Let us close this chapter with a quotation of a similar point made by Mook (1983), "… psychological investigations are accused of 'failure to generalize to the real world' because of sample bias or artificiality of setting … such 'generalizations' often are not intended. Rather than making predictions about the real world from the laboratory, we may test predictions that specify what ought to happen in the lab. We may regard even 'artificial' findings as interesting because they show what can occur," [p. 379], even when the precise conditions in which they actually occur are not (yet) known.

References

Bar-Hillel, M. (1984). Representativeness and fallacies of probability judgment. *Acta Psychologica*, *55*(2), 91–107. https://doi.org/10.1016/0001-6918(84)90062-3

Bernoulli, J. (1713). *Ars conjectandi, opus posthumum* [The art of conjecturing, posthumous work]. Basel, Switzerland: Thurneysen Brothers.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe, A. P. Shimamura, J. Metcalfe, A. P. Shimamura (Eds.) , *Metacognition: Knowing about knowing* (pp. 185-205). Cambridge, MA US: The MIT Press.

Bodemer, N., & Gaissmaier, W. (2012). Risk communication in health. In *Handbook of risk theory* (pp. 621-660). Springer, Dordrecht.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., ... & Altmejd, A. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behavior*, *2*(9), 637.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*(4), 297-312.

Campbell, D.T. (1996). Regression artifacts in time-series and longitudinal data. *Evaluation and Program Planning, 19*, 377-389.

Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts*. New York, NY, US: Guilford Press.

Cialdini, R. B. (2012). The focus theory of normative conduct. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology., Vol. 2.* (pp. 295–312). Thousand Oaks, CA: Sage Publications Ltd. https://doi.org/10.4135/9781446249222.n41

Cooper, J. (2012). Cognitive dissonance theory. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology., Vol. 1.* (pp. 377–397). Thousand Oaks, CA: Sage Publications Ltd. https://doi.org/10.4135/9781446249215.n19

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over-and underconfidence: The role of error in judgment processes. *Psychological Review*, *101*(3), 519-527. doi:10.1037/0033-295X.101.3.519

Fiedler, K. (2008). The ultimate sampling dilemma in experience-based decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(1), 186–203. https://doi.org/10.1037/0278-7393.34.1.186

Fiedler, K. (2010). The asymmetry of causal and diagnostic inferences: A challenge for the study of implicit attitudes. In J. P. Forgas, J. Cooper, & W. D. Crano (Eds.), *The psychology of attitudes and attitude change.* (pp. 75–92). New York, NY: Psychology Press.

Fiedler, K., Brinkmann, B., Betsch, T., & Wild, B. (2000). A sampling approach to biases in conditional probability judgments: Beyond base rate neglect and statistical format. *Journal of Experimental Psychology: General*, *129*, 399-418.

Fiedler, K., Hofferbert, J., Woellert, F., Krüger, T., & Koch, A. (2015). The tragedy of democratic decision making. In J. P. Forgas, K. Fiedler, & W. D. Crano (Eds.), *Social psychology and politics.* (Vol. 17, pp. 193–208). New York, NY: Psychology Press.

Fiedler, K., Hütter, M., Schott, M., & Kutzner, F. (2019). Metacognitive myopia and the overutilization of misleading advice. *Journal of Behavioral Decision Making*. https://doi.org/10.1002/bdm.2109

Fiedler, K., & Krueger, J. I. (2012). More than an artifact: Regression as a theoretical construct. In J. I. Krueger (Ed.), *Social judgment and decision making* (pp. 171–189). New York, NY: Psychology Press.

Fiedler, K., & Kutzner, F. (2015). Information sampling and reasoning biases: Implications for research in judgment and decision making. In G. Keren and G. Wu (Eds.), *The Wiley Blackwell Handbook of Judgment and Decision Making* (pp. 380-403). New York: Wiley.

Fiedler, K., & Prager, J. (2018). The regression trap and other pitfalls of replication science—Illustrated by the report of the Open Science Collaboration. *Basic and Applied Social Psychology*, *40*(3), 115–124. https://doi.org/10.1080/01973533.2017.1421953

Fiedler, K., Schmid, J., & Stahl, T. (2002). What is the current truth about polygraph lie detection. *Basic and Applied Social Psychology*, *24*(4), 313–324. https://doi.org/10.1207/S15324834BASP2404_6

Fiedler, K., & Unkelbach, C. (2014). Regressive judgment: Implications of a universal property of the empirical world. *Current Directions in Psychological Science*, *23*(5), 361-367.

Forgas, J. P., & Ciarrochi, J. V. (2002). On managing moods: Evidence for the role of homeostatic cognitive strategies in affect regulation. *Personality and Social Psychology Bulletin*, *28*(3), 336–345. https://doi.org/10.1177/0146167202286005

Galton, F. (1907). Vox populi. *Nature*, *75*, 450–451.

Gavanski, I., & Hui, C. (1992). Natural sample spaces and uncertain belief. *Journal of Personality and Social Psychology*, *63*(5), 766–780. https://doi.org/10.1037/0022-3514.63.5.766

Heck, D. W., Thielmann, I., Klein, S. A., & Hilbig, B. E. (2019). *On the limited generality of air pollution and anxiety as causal determinants of unethical behavior*. Manuscript under review.

Higgins, E. T. (2012). Regulatory focus theory. In P. M. Van Lange, A. W. Kruglanski, E. T. Higgins, P. M. Van Lange, A. W. Kruglanski, E. T. Higgins (Eds.) , *Handbook of theories of social psychology (Vol 1)* (pp. 483-504). Thousand Oaks, CA: Sage Publications Ltd.

Hogarth, R. M., & Karelaia, N. (2005). Ignoring information in binary choice with continuous variables: When is less "more"?. *Journal of Mathematical Psychology*, *49*(2), 115-124.

Ioannidis, J. P. (2005). Why most published research findings are false. *Chance*, *18*(4), 40-47.

Johnson, E. J., & Goldstein, D. G. (2003). Do defaults save lives? *Science, 302,* 1338–1339.

Juslin, P., Winman, A., & Hansson, P. (2007). The naïve intuitive statistician: A naïve sampling model of intuitive confidence intervals. *Psychological Review*, *114*(3), 678-703.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction?" *Psychological Science*, *19*(6), 585–592. https://doi.org/10.1111/j.1467-9280.2008.02127.x

Lewin, K. (1943). Psychology and the process of group living. *The Journal of Social Psychology, SPSSI Bulletin, 17*, 113–131.

Lu, J. G., Lee, J. J., Gino, F., & Galinsky, A. D. (2018). Polluted morality: Air pollution predicts criminal activity and unethical behavior. *Psychological Science*, *29*(3), 340–355. https://doi.org/10.1177/0956797617735807

Metcalfe, J., & Kornell, N. (2007). Principles of cognitive science in education: The effects of generation, errors, and feedback. *Psychonomic Bulletin & Review*, *14*(2), 225–229. https://doi.org/10.3758/BF03194056

Mook, D.G. (1983). In defense of external invalidity. *American* Psychologist, 38(4), 379-387.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502-517. doi:10.1037/0033-295X.115.2.502

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

Prager, J., Krueger, J. I., & Fiedler, K. (2018). Towards a deeper understanding of impression formation—New insights gained from a cognitive-ecological perspective. *Journal of Personality and Social Psychology*, *115*(3), 379–397. https://doi.org/10.1037/pspa0000123

Ritov, I. (1996). Anchoring in simulated competitive market negotiation. *Organizational Behavior and Human Decision Processes*, *67*(1), 16–25. https://doi.org/10.1006/obhd.1996.0062

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, *15*, 351–357. https://doi.org/10.2307/2087176

Sorkin, R. D., Hays, C. J., & West, R. (2001). Signal-detection analysis of group decision making. *Psychological Review*, *108*(1), 183–203. https://doi.org/10.1037/0033-295X.108.1.183

Stelzl, I. (1982). *Fehler und Fallen der Statistik*. [*Errors and Traps of Statistics*]. Bern: Huber.

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. New York, NY, US: Doubleday & Co.

Swets, J., Dawes, R.M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*, Whole No. 1.

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review*, *117*(2), 440–463.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Vogel, T., Kutzner, F., Fiedler, K., & Freytag, P. (2013). How majority members become associated with rare attributes: Ecological correlations in stereotype formation. *Social Cognition*, *31*(4), 427–442. https://doi.org/10.1521/soco_2012_1002

Wells, G. L., Malpass, R. S., Lindsay, R. C., Fisher, R. P., Turtle, J. W., & Fulero, S. M. (2000). From the lab to the police station: A successful application of eyewitness research. *American Psychologist*, *55*(6), 581-598.

Wells, G. L., & Elizabeth Luus, C. A. (1990). Police lineups as experiments: Social methodology as a framework for properly conducted lineups. *Personality and Social Psychology Bulletin*, *16*(1), 106-117.

Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest*, *18*(1), 10–65. https://doi.org/10.1177/1529100616686966