Grażyna Wieczorkowska, Grzegorz Król

The Pittfalls of Research Practices in Management Science

Problemy Zarządzania 14/2 (2), 173-187

2016

Problemy Zarządzania, vol. 14, nr 2 (60), t. 2: 173–187
ISSN 1644-9584, © Wydział Zarządzania UW
DOI 10.7172/1644-9584.60.11

Ten Pitfalls of Research Practices in Management Science

Submitted: 10.07.16 | Accepted: 18.08.16

Grażyna Wieczorkowska*, Grzegorz Król**

This essay-like text discusses 10 pitfalls in the analysis of data in management science: (1) Too small a number of experimental studies; (2) Ignoring the specifics of the research object; (3) Lack of standard operationalizations; (4) Weakness of measurement ; (5) Weakness of data analyses; (6) Too high a level of generality of theory and too few replications; (7) Misinterpretation of the outcomes of statistical analyses; (8) Reviewers' expectation regarding the samples and hypotheses testing; (9) Missing "time" in the list of predictors; (10) Wrong standards of publication. Most of these risks apply also to psychological and sociological research.

Keywords: research practices, research pitfalls, data analysis, management science.

Dziesięć pułapek badawczych w naukach o zarządzaniu

Nadesłany: 10.07.16 | Zaakceptowany do druku: 18.08.16

W niniejszym tekście omawiamy 10 pułapek analiz danych w naukach o zarządzaniu: (1) zbyt mała liczba badań doświadczalnych; (2) ignorowanie specyfiki przedmiotu badań; (3) brak standardowych operacjonalizacji; (4) słabość pomiaru; (5) słabość analiz danych; (6) zbyt wysoki poziom ogólności teorii i zbyt mała liczba replikacji; (7) błędna interpretacja wyników analiz statystycznych; (8) oczekiwania recenzentów dotyczące doboru próby i testowania hipotez; (9) brak „czasu" na liście predyktorów; (10) niewłaściwe standardy publikacji. Większość z tych zjawisk odnosi się również do badań psychologicznych i socjologicznych.

Słowa kluczowe: praktyki badawcze, pułapki analiz, analizy danych, nauki o zarządzaniu.

JEL: C18

* **Grażyna Wieczorkowska** – prof. dr hab., Faculty of Management, University of Warsaw.

** **Grzegorz Król** – dr, Faculty of Management, University of Warsaw.

Correspondence address: Faculty of Management, University of Warsaw, Szturmowa St. 1/3, 02-678 Warsaw, Poland; e-mail: gw@uw.edu.pl.

Ministry of Science and Higher Education
Republic of Poland

The creation of the English-language version of these publications is financed in the framework of contract No. 768/P-DUN/2016 by the Ministry of Science and Higher Education committed to activities aimed at the promotion of education.

Introduction from Grażyna Wieczorkowska

For more than 35 years I have spent many hours a week analyzing research results, overseeing analyzes carried out by my doctoral students, or writing reviews of doctoral theses, dissertations and scientific articles in psychology, sociology and management.

In decades past, statistical analyzes were usually performed by specialists, but today's statistical packages allow even complex analyses to be run by anyone in a few minutes.

While the shift has made many researchers more productive, it has not always made them more reflective of the research practices – from research design to conclusions, nor has it made them more prone to avoid several analytical pitfalls. Easy execution of advanced statistical analyses is rarely combined with an appropriate level of methodological concern.

Nine years ago, when I switched faculties from psychology to management science, I began to review management research. My previous work in social psychology taught me to look carefully for methodological and analytical pitfalls that might reduce the validity of the research I reviewed. Curiously, I found almost no discussion of these pitfalls in the management research methods literature. This prompted me to collaborate with my colleague, Dr. Grzegorz Król, in writing the present paper. We chose most frequent – in our opinion – pitfalls in the management research we see. Each of these ten is discussed below, leading to our recommendations for improving the quality of management research.

1. Too small a number of experimental studies

Our knowledge is built mainly by results of experimental studies; correlational research rarely plays such an important role. In contrast, experiments are still rare in the management sciences, where correlational studies abound. The development of behavioral economics recognized by two Nobel Prizes in economics in 2002 for: (1) Vernon Smith, for the establishment of laboratory experiments as a tool for empirical analysis, and (2) Daniel Kahneman, for the use of psychological tools in economic research; stimulated the growth of laboratory experiments to test economic ideas. A similar growth has not yet occurred in management science.

Experimental study differs from other types of scientific inquiry; instead of waiting for the natural occurrence of events of interest as they happen, experiments create the conditions required for observation (more: Aronson and Wieczorkowska, 2001).

Critics of the experimental approach often note that most laboratory experiments employ only small numbers of undergraduate students, and thus have questionable external validity. Experiments allowing random assignment to experimental conditions create "collective clones"; that protects

 DOI 10.7172/1644-9584.60.11

research from the distorting influences of other, confounding variables. As a result, experiments with small samples of undergraduates can frequently untangle causal connections that are impossible to assess in correlational studies, including questionnaire studies involving thousands of managers. Laboratory experiments are sometimes dismissed as unrealistic imitations of human interaction, unrepresentative of the real world. Those who make such claims often forget that an experiment can be realistic in two ways (more: Aronson and Wieczorkowska, 2001). *Situational realism* occurs when the situation in the laboratory is similar to those people often experience in the outside world. *Psychological realism* occurs when the laboratory situation encourages participants to treat the research situation seriously. Good research tasks, such as those in which participants take part in an involving game, engage participants and cause the situation to become very real. The feedback participants receive has a real value. Participants in these research tasks are much more involved than they are completing survey questionnaires.

Experimental methods are sometimes questioned on ethical grounds. A very few drastic side-effects of participation, like emotional stress of the subjects in Milgram (1963) experiment, which facilitated, in some of the subjects, learning something potentially unpleasant about themselves, resulted in developing VERY strict ethic research rules. In our opinion too strict. For example, the experiment (Piliavin at al., 1969) in which the helping behavior of unsuspecting passengers on a subway train was measured has shown that even generally very helpful people are more reluctant to help a drunk. If the participants possessed prior information, and knew that they were being watched, the bystanders would be more likely to help. The usefulness of the research results is unquestionable, but the fact that there was no pre-experimental consent means that this experiment would not be allowed today. The pendulum of our ethic concerns swung too far the other way.

1.1. Two-sentence summary

Our knowledge is built mainly by results of experimental studies, to a much lesser extent by results of correlational research. Psychological realism of experimental situation is much more important than the situational one.

2. Ignoring the specifics of the research object

Much of the social science methodology is modeled on research methods of biology and physics, but **the object of our study: human behavior and the minds that control it** is much more **complicated** and **reactive** . Perhaps that is why Taleb (2012) sarcastically compares the attempt of social sciences to apply the methodology used by physicists to cows that attempt to

fly. Consider, for example, the results of a **"relaxation training"** exercise I (GW) conducted with students during a class on *Psychosomatic Medicine.* Almost 300 students present in the classroom were asked to follow the instructions, while relaxing music played in the background. I watched the reaction of the participants, and asked them to rate their relaxation "success" by selecting one of 6 categories. Most students selected options indicating a degree of success in this task. But 8% chose the answer "**I did not want to relax**, so I have not tried," and 9% showed a contrasting answer: "**Trying to relax irritated me**". Such responses are contrary to the assumption that the introduction of music would simply push all respondents up a notch the relaxation variable.

Variance in mental response tends to be much larger than variances in physical or biological responses. An administration of sedatives to the participants would not invoke a contrast effect – all subjects will fall asleep sooner or later. An administration of a relaxation procedure would evoke a contrast effect in some participants – some not only failed to relax, but they presented the opposite reaction – irritation. As a result, our models ignoring individual differences and situational context cannot explain large proportions of response variance. By focusing our statistical analyses on measures of central tendency, we often fail to see contrarian effects – in this example, instead of lack of relaxation, the opposite effect: tension.

2.1. Two-sentence summary

Reactions of human mind are more variable than reactions of human body which in turn is more variable than the reaction of inanimate objects. We can not only fail in getting a desired effect of our stimulus, but instead get a counter-effect.

3. Lack of standard operationalizations

A weakness of the social sciences is the **lack of systematization** of concepts. There is too much creativity: continuous introduction of new concepts, too few replicable findings.

It is no secret that the easiest way to achieve high citation frequency is by introducing a new concept and publishing a questionnaire to measure it. If we are lucky, many people will begin to correlate the new scale with existing ones until a wave of interest subsides.

We often try to avoid using the science "jargon" in professional journal publications, in part to attract the interest of non-professionals.

Motivation, trust, leadership style and other concepts we study have **prototypical structure and fuzzy borders**, so their definition in a classic way (by giving necessary and sufficient conditions) is not possible. This is why **operationalizations** of our concepts are so important. Unfortunately, because of specificity of our object, there are no standard operationaliza-

 DOI 10.7172/1644-9584.60.11

tions, contrary to e.g. operationalization of inflation in economics, anemia in medicine etc.

Operationalization in experimental studies needs a description of the way we can *manipulate (change values of) independent variable, e.g "level of threat"*.

Operationalization in correlational studies needs a description of the way we can measure variables. If we would like to study the effect of threat on negotiation outcomes, we have to take into account that threat and negotiation outcomes might be operationalized quite differently from one organization or culture to the next.

The challenges of **context-dependent meaning** can be found in standardized tests as well. Many questionnaires showing good psychometric properties of reliability and validity in one language and culture do not survive translation or transport. For example, the meaning of a questionnaire item such as "I am nonchalant about details" can change dramatically when read by pharmacists or psychologists, surgeons or artists, programmers or company presidents. As a result, we need to test our off-the-shelf research instruments on local samples to see if we are measuring what we assume the tools are measuring.

3.1. Two-sentence summary

Our object of study does not allow standard operationalization but technology development can solve this problem in the future. Psychometric values of questionnaire scales reported in literature could be different on our sample.

4. Weakness of measurement

A well-known saying by Gordon Allport: **"If we want to know how people feel – why not ask them**?" marked the beginning of popularity of self-reporting techniques to measure various characteristics. By creating various tools based on self-description we forget that it was shown in a series of studies (more: Aronson and Wieczorkowska, 2001) that we are often unaware of influences which we are subject to, yet that does not prevent us from believing that we can accurately identify the causes of our thoughts, feelings and behaviors. For some people the answers to the questionnaire items are the result of reflection, others might not have prior thoughts on a given subject, so their answers are created on the fly. An answer to a single question may have a large **measurement error** – therefore in physics, chemistry or biology, many repetitions of the measurement are made. Unfortunately, it is possible only in the study of inanimate beings, more precisely beings without memory and free will. People remember that they were just asked a certain question, and repeating it will cause their irritation.

Therefore, instead of repeating the measurement, we ask several similar questions. To prove the consistency of responses to all items we compute **Cronbach's** alpha – the most popular measure of reliability (homogeneity) of the synthetic index. Let's take an example of a very popular measurement of **Five-Factor Model of Personality** (Costa and McCrae, 1992), based on lexical research. An abridged version of NEO_FFI **(NEO Five-Factor Inventory)** consists of 60 questions that describe a respondent in 5 dimensions: neuroticism, extraversion, openness to experience, agreeableness and conscientiousness, for which Cronbach's alpha for a group of over 1,000 students (Turska, 2014) amounted to 0.842; 0.704; 0.69; 0.807; 0.712. None of these 12-item indices are univariate. Extraversion and openness to experience consist of as many as 4 factors each. This is not surprising when we look at the content of the questions. One source of multidimensionality is caused by different cognitive processes activated when providing consent and disagreement (see e.g. research on asymmetry of Wanke at al., 1995), often causing positive (requiring consent) and negative (requiring disagreement) items in the factor analysis to be separated into different factors. The biggest problem is the **heterogeneity of theoretical** constructs. In the *openness to experience* index, up to a quarter of the questions refer to an interest in art / poetry. The conscientiousness scale contains the need for achievement, responsibility, and meticulousness. This heterogeneity makes it difficult to imagine a person who has acquired a high / low score on the scale. The problem was already pointed out by Allport (1940) claiming that results obtained in the factor analysis are averaged dimensions of personality, which is a **total abstraction**, unsuitable for the psychologist who wants to explore the personality of individuals. We would soften this claim, saying that measuring the properties of a person in this way causes **a gap between** the result of **measurement** and **observation.** We analyze **statistical abstractions.** Cronbach's alpha value is treated as the most important measure of psychometric value of the indicator, because we keep forgetting how easy it is to obtain a high reliability alpha coefficient. It is enough to ask the same question worded in different ways.

Do components of a synthetic index have to be highly correlated with each other? In some cases we can talk about a **trade-off between validity and reliability**. Searching for an indicator of the risk of being overweight, we can ask about the frequency of eating sweets, drinking beer, eating at night... Each of these activities can lead to weight gain, acting alone. Another example: a synthetic indicator of economic activity (Czapiński, 1996) constructed with the ratings of 8 possible manifestations of the latent variable, e.g. whether they invest in services, in trade, in stocks, whether they increase their skills, have non-professional activities, have plans for the future etc. Each item describes a specific manifestation of the latent construct "economic activity", but does have to correlate with others. The resulting Cronbach's alpha for the index may be low, but predictive validity of this index turned out to be high.

 DOI 10.7172/1644-9584.60.11

Therefore, we should not – even though it is commonly done – identify the quality of a synthetic index with the value of Cronbach's alpha. Students keep asking "How high should be the value of Cronbach's alpha?" We should keep in mind that even a high alpha **does not warrant** either **unidimensionality** or the construct's **validity**. Unfortunately, in assessing the quality of synthetic indices built from multiple questions, we limit ourselves too often to testing reliability, not reflecting on the validity of the measurement.

In our opinion, it appears that the development of questionnaires looks like a road with a dead end, as evidenced by the **lack of substantial progress in psychometrics** over the past several decades. However, we have to keep in mind that an **imperfect measurement tool is better than the lack of tools** – such as using bad sewing machine is better than hand-sewing a suit (Gilbert, 2007). The imperfection of the measurement is a problem that disqualifies our results only when we do not see it. If we are aware of the inevitable distortions, which are subject to our self-reports, we seek to make appropriate adjustments. Let's hope that technology development, which we await impatiently, can solve soon many measurement problems.

4.1. Two-sentence summary

Researchers too often focus on measurement reliability and overlook its validity. High Cronbach's alpha does not guarantee either unidimensionality or the construct validity.

5. Weakness of data analyses

In **the physical sciences or economics,** the entities measured are **real** (e.g. weight, length of an object; amount in €). Unit meter, for example, has got a standard against which we can compare our measure. In **social sciences**, the subjects of our measurements are **virtual entities** (theoretical constructs, e.g. cognitive representations, attitudes, traits, affective states) and we do not have a standard unit for attitude, or intelligence, against which we can compare the results of our measurement. An economist is interested in the amount of money available on the market – so he will try to objectively measure the level of earnings. In social studies, even if we ask our respondents to specify their earnings in PLN, EUR …., we are most interested in the cognitive representation of earnings in the mind of the respondent. Therefore, we do not take as seriously as researchers in the physical sciences the original units of measurement, frequently performing transformations resulting in a change of the units, e.g. converting earnings into a log measure to reduce the skewness of the distribution. **Methodologically rigorous researchers** claim that the rating scales, e.g. when asking about the level of happiness, are not quantitative, because they do not have a fixed unit of measurement: e.g. the distances between the points at the

rating scale "3 = rather happy" and "4 = happy", and "2 = rather unhappy" are not the same, although it is a 1-point difference. The same researchers, however, forget about the requirements of the stability of the unit of measurement when they calculate the average of school grades. There is no evidence that the differences between the assessments of "4 = good", " 5 = very good", and "3 = sufficient" are the same.

This is what we called methodological hypocrisy – the gap between what is said in statistical handbooks and what is done in practice. For sure, the results of a statistical analysis may be distorted by the lack of fulfillment of the assumptions (e.g. normality). But more important is comparability of results collected in different laboratories. If everybody applies parametric tests, and we use a non-parametric test – although we follow the rules – this will prevent the comparison. It is, however, more important that in all comparative studies the distortion, e.g. related to the measurement scale, was the same. Especially funny is the situation when the rigorously rooted researcher uses non-parametric tests of significance for difference in means, and later uses the same data for structural modeling (e.g. Turska, 2014).

While experimental studies have good and clear standards of analysis, in correlational research, we publish a lot of results of dubious value (e.g. Brzeziński, 2012; Starbuck, 2016). Eminent psychologist prof. Robert Nisbett (2016) even announces **"The Crusade Against Multiple Regression Analysis"**. The problem stems mainly from low internal validity which is an inherent property of correlational studies and is manifested in confounding influence of various variables. **Multivariate analysis of correlational data is a must**, but the problem lies in its **sensitivity to the validity of the model**. As we repeat unendingly to our students: "regression coefficients depend on the company". Only those not understanding the essence of these analyzes can decide to use a stepwise regression analysis (leaving substantive decisions to statistical algorithms), and write with full conviction that a factor analysis or a cluster analysis proved an existence of k factors or clusters. Too often, removing one variable changes the structure of the correlation matrix. Our students are impressed when we show them an example of analysis (Wieczorkowska, Wierzbiński, 2011) in which, using the same data, the relationship between X and Y changes, depending on which additional variables are controlled. Of course, where there is a clear theory, the list of variables is uniquely determined. In the analyzes of the survey data on representative samples, where controlling sociodemographic variables is a necessity, we often forget about entering interactive effects into the model. For example, checking the sociodemographic predictors of the level of education (measured in years), we find no significant relationship with gender, because a significant predictor is the interaction of age and gender. The educational advantage of older men is matched with the educational advantage of younger women; as a result, the main effect of gender is not significant. A similar relationship was found in 14 of the 33

 DOI 10.7172/1644-9584.60.11

countries studied (Wierzbiński, 2009). It required introduction of interactive effects in the regression model, which researchers often ignore.

It is true, however, that many issues cannot be tested experimentally and we should intensively work on developing a new model of analysis of correlational data. A simple alternative like Ordinal Pattern Analysis (Thorngate et al., 2016) is an example. The task is the more important, as there are huge collections of publicly available data on representative samples, **collected with an enormous effort and cost**, but cognitively rather **poorly exploited**.

5.1. Two-sentence summary

Methodological purity often makes our laboratory work sterile and fruitless. Methodological "dirt" is introduced by flawed measurement, so meeting rigorous statistical assumptions does not guarantee validity of the statistical analyses.

6. Too high a level of generality of theory and too few replications

We keep forgetting that the results of our research are highly context dependent. Relationships established in the „ceteris paribus" model are very susceptible to the „third variable" effect. We envy physicists and we would like to develop such simple rules as Newton's three principles of dynamics, thus ignoring the fact that our mind develops in response to environmental challenges, and is therefore a collection of modules of learned response patterns to various classes (!) of stimuli. A gap between research in **nomothetic** research and practice, which requires **idiographic** knowledge, is also a serious problem**.** A big progress in our scientific knowledge does not correspond to the progress in the efficiency of practitioners. There is no easy way to convert our scientific expertise into practical one.

It is more useful to aim at describing a model of job interview for e.g. IT Executive Interview, test it, then think about a model for job interview for elementary school teachers, rather than aim from the start at creating a General Model of Job Interview.

Our theories are losing their explanatory values – when I (GW) started my scientific work over 35 years ago, speaking about unconscious processes was totally rejected by science; today there is no doubt that the processing of information takes place mostly outside of consciousness. Empirical data stays useful longer than theories. Still we have too few replications – because hardly any successful replication in the social sciences can be considered trivial (more: Aronson, Wieczorkowska, 2001). In recent years, we have had to deal with a crisis in science in general, with the question of replicability. It shows how multivariate the matter we have to deal with is (e.g. van Babel et al., 2016). Such replication failures should be an impetus to the

search for the boundary conditions for the previously described relationships. Information on "when it does not work" can extend our knowledge.

6.1. Two-sentence summary

Social sciences are context dependent, so the model we explore should not be formulated on too high a level of abstraction. We should not forget that our findings are always embedded in some cultural, economic and social context.

7. Misinterpretation of the outcomes of statistical analyses

Rigorous methodological expert William Starbuck (2016, p. 171), in his very interesting essay (though we do not agree with all his points) criticized the two common (unfortunately) success-facilitating practices:

> **"HARKers**[1] gather data first, make statistical analyzes, then formulate hypotheses, and finally search for theories or previous studies that support or contradict the newly invented hypotheses ... Data mining, **p-hacking,** or data dredging involves subjecting data to many calculations or manipulations in search of an equation or classification system that captures strong patterns."

We do not know why it is not widely accepted that a regression equation should be developed on one half of the available data, and tested (Pedhazur, 1997) on the other half. We need new models of analyzing large data sets. In our opinion, **the fashion for structural modeling** brings far more harm than good, because it results in numerous publications adding very little to our knowledge: a model can always be found to fit the data, and the researcher's mind is always able to adapt a new theory to the data.

The economists are proud of their high fits in their analyses and they disavow the results of human resources specialists who can explain a very small percentage of their dependent variables. They keep forgetting about the difference between individual indicators (describing the respondent) and group indicators (describing complex, aggregate objects, e.g. country). For example (Wieczorkowska, Król and Wierzbiński, 2015), in the European Social Survey[2] respondents answer the question "How happy are you?", using an 11-point response scale. While the distribution of the respondent-level indicator of happiness has a normal-like shape, the distribution of the group-level indicator (national average for the country) is uniform – no two countries have the same average. While the respondent-level indicator can be regarded as a discrete variable, the group-level indicator is a continuous variable. As a consequence, different results are obtained, when testing the same hypothesis on the relationship between e.g. left-right wing political attitude and happiness (Napier and Jost, 2008), on the country versus respondent-level indicators. The percentage of explained variance is incomparably smaller when we analyze the individual-level indicators than the aggregate-level ones, because the latter have much smaller variance.

 DOI 10.7172/1644-9584.60.11

7.1. Two-sentence summary

We need new models for analyzing large data sets. One simple heuristic: use half of our data to derive models of relationships between variables, then test the models with the other half.

8. Reviewers' expectation regarding the samples and hypotheses testing

Very frequent objections raised by the reviewers in management science are associated with (1) **the lack of representativeness of the sample**, (2) **the lack of hypotheses.** It is often ignored that representative samples are necessary if the aim of the researcher is the **estimation of distributions** of variables in the population – e.g. if we want to predict the results of the elections (but again, we should look for a representative sample of those who really go voting, and not those who are entitled to vote). It is easy to have a representative sample of inanimate objects, for example screwdrivers in a factory line, because they **cannot refuse to participate in the study.** When the studied object are people, we can draw a sample to meet certain criteria, but **we cannot guarantee** that selected people will want to participate in the study. The survey response rate has lowered in recent decades twice. To make matters worse, we have to deal with **false respondents** – who agree to participate in the survey, but provide their answers randomly. Therefore, there is nothing wrong in the fact that the hypotheses testing in most of the studies uses convenience samples. It would be nice to be able to demonstrate the external validity (the ability to generalize results to the population) of the results, but the internal validity (first you have to have something to generalize) is much more important. Rather than trying to get representative samples, a much better strategy to maximize the external validity are replications of research trials on homogeneous samples – separately investigating farmers, separately academics, separately students etc.

The second objection comes from the false belief that scientific work must always be carried out under the deductive approach, which – in particular in the management – is not true. Some work in management sciences are attempts to synthesize knowledge, to create a model that is not prepared to be tested empirically. In such a case, adding hypotheses could look only funny.

8.1. Two-sentence summary

Random, representative sample can only be achieved when the research participants have no possibility of excluding themselves (behaviorally or mentally) from being examined. Replicating findings with many convenience samples is more useful.

9. Missing 'time' in the list of predictors

Our object of study is capable not only of refusing to participate in the research, getting irritated with it, or – which happens often – trying to influence the outcomes, but also, unfortunately, it has a memory.

Asking the same question multiple times causes difficulties of interpretation, because successive measurements are not independent of each other. Therefore, the vast majority of both experimental and correlational research deals with a single time point. As a result, we are not able to capture the dynamics of the processes. Even if we show a positive effect of drinking a cup of coffee on a test score, we will not know what happens several hours later, when the subjects have left the laboratory. Even if nothing in the environment has changed, the repetition of a stimulus modifies reactions. These changes are described by three effects: mere exposure, habituation and oversaturation. The **mere exposure effect** comes into play when initially neutral objects are evaluated the more positively the more often they are exposed. Hence the desire of politicians to increase their exposure – yet they forget, that if the initial reaction was negative, an exposure will magnify the negative reaction. If we ask for a second time, or a third, for the respondent's attitude to tattoos, we risk that even if the initial assessment showed indifference, we may be getting more and more positive assessments in subsequent measurements. If we verify experimentally an impact of praises from the superior on the motivation to work, we will not grasp the impact of the second, third, fiftieth praise. Psychology shows that each further praise will have a weaker rewarding value – other things being equal. **Positive emotions usually subside with frequent contact with the object that inspires them** – we quickly get used to the good. This phenomenon is called **hedonistic habituation**. Even the greatest praise loses its value when we hear it too often. At some point, **saturation** takes place, and later **oversaturation**. What fascinated us at the beginning, begins to irritate us. Although we know that many interesting phenomena, such as motivation, happiness, have **wave-like qualities**, we will not be able to monitor their dynamics as long as we cannot conduct the measurement without keeping our subjects unaware of the process. Fortunately, new technologies bring hopes for new developments in this area.

9.1. Two-sentence summary

The main challenge in social science is to study the dynamics of previously-described relationships. Even if nothing in the environment has changed, the repetition of a stimulus could result in any of three effects: mere exposure, habituation or oversaturation.

 DOI 10.7172/1644-9584.60.11

10. Wrong standards of publication

We have to deal with uncontrollable flood of publications – scholars are forced to publish whether they have something important to say or not (Wieczorkowska, Król and Wierzbiński, 2015). The increasing number of research papers exceeds the possibility of integration, unless changes to the **standards of publication** take place – standards which have not changed significantly since the pre-internet era. Extracting the key ideas from the growing avalanche of texts that we read requires us to ignore a lot of information, e.g. the names of the authors in-lined into the text, and this process reduces the reader's cognitive resources (Gilbert et al., 1993). There are **too many references** in our publications, not adding anything to the line of reasoning. With a current quantum leap in the number of publications, reporting a history of research on a given issue should be left to historians of science. More attention should be focused on solving the problem than on reporting a history of previous attempts.

Division of each publication into the substantive and technical part – the latter made available on the internet – would allow for obligatory making available a file with raw data, which in turn would allow those interested to attempt a much better integration than the **meta-analysis approach** invented in the pre-internet era.

10.1. Two-sentence summary

Being busy fulfilling our old publication standards, we ignore the changes that have been taking place in our information environment. Reporting a history of research on a given issue should be left to historians of science.

Conclusions

The biggest challenge – in our opinion – is underestimating the weakness of measurement and fetish-like treatment of statistical significance. Even the best analysis does not help if the measurement is not valid. We agree with Taleb (2012), who in this last book "Antifragile: Things That Gain from Disorder" illustrated the problem nicely by showing the relation between the number of variables and number of spurious correlations in Figure 18 titled "The Tragedy of Big Data". **The more variables, the more correlations that can show significance** in the hands of a "skilled" researcher.

Researchers forget that the main goal is to understand our data; statistical significance is only a stamp that we add at the end. The diagnosis of the causes made by Starbuck (2016, p. 171) is depressing, but it is hard to disagree with it (Wieczorkowska, Król & Wierzbiński, 2015).

> „**Academic culture has become cynical and careerist,** in part because universities use characteristics of research publications when they evaluate faculty or advertise faculty achievements. Professors want to keep their jobs and to attain promotions.

Universities want to claim that their faculty members have made "significant" contributions. Therefore, there is unremitting pressure to lower the criteria for "significant findings" to levels that every researcher and every study can meet".

Endnote

1 HARKing (Hypothesizing After Results are Known).

2 (str 182 w 12 dół jest odnośnik 2 czy mam wyrzućić?)

References

Allport, G.W. (1937). *Personality: A psychological interpretation*. New York: Holt.

Aronson, E. and Wieczorkowska, G. (2001). *Kontrola naszych myśli i uczuć. Rozdział 4. Jak możemy zdobyć odpowiedź na nurtujące nas pytania?* Santorski: Warszawa.

Brzeziński, J.M. (2012). Co to znaczy, że wyniki przeprowadzonych przez psychologów badań naukowych poddawane są analizie statystycznej? *Roczniki Psychologiczne, 15*(3), 7–40.

Costa, P.T. and McCrae, R.R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five- -Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resourses.

Czapiński, J. (1996). Uziemienie polskiej duszy. In: M. Marody and E. Gucwa-Leśny (eds), *Podstawy życia społecznego w Polsce* (pp. 252–275). Warszawa: Wydawnictwo Instytutu Studiów Społecznych Uniwersytetu Warszawskiego.

Davison, A.C and Hinkley, D.V. (1997). Bootstrap methods and their application. *Cambridge Series in Statistical and Probabilistic Mathematics*. New York: Cambridge University Press.

Gilbert, D. (2007). *Stumbling on Happiness*. New York, NY: Vintage Books.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, 67, 371–378.

Nisbett, R. (2016). https://www.edge.org/conversation/richard_nisbett-the-crusade-against-multiple-regression-analysis

Pedhazur, E.J. (1997). *Multiple regression in behavioral research*. Belmont: Wadsworth Publishing.

Piliavin, I.M., Rodin, J. and Piliavin, J. (1969). Good Samaritanism: an underground phenomenon? *Journal of Personality and Social Psychology*, *13*(4), 289–299.

Starbuck, W.H. (2016). 60th Anniversary Essay: How Journals Could Improve Research Practices in Social Science. *Administrative Science Quarterly*, *61*(2), 165–183. http://doi.org/10.1177/0001839216629644

Taleb, N. (2012). *Antifragile: Things That Gain from Disorder*. New York: Random House.

Thorngate, W. and Chunyun, M. (2016). Wiggles and Curves: The Analysis of Ordinal Patterns. *Problemy Zarządzania*, 2.

Turska E. (2014). *Kapitał kariery ludzi młodych. Uwarunkowania i konsekwencje*. Katowice: Wydawnictwo Uniwersytetu Śląskiego.

Van Bavel, J.J., Mende-Siedlecki, P., Brady, W.J and Reinero, D.A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences*. Social Sciences – PNAS, 23, 6454-6459, doi: 10.1073/pnas.1521897113.

Wänke, M., Schwarz, N. and Noelle-Neumann, E. (1995). Asking comparative questions: The impact of the direction of comparison. *Public Opinion Quarterly*, 59, 347–372.

 DOI 10.7172/1644-9584.60.11

Wierzbiński, J. (2009). *Badanie zaufania do organizacji: problemy metodologiczne*. Warszawa: Wydawnictwo Naukowe Wydziału Zarządzania UW.
Wieczorkowska, G. and Wierzbiński, J. (2011). *Statystyka: od teorii do praktyki*. Warszawa: Wydawnictwo Naukowe Scholar.
Wieczorkowska, G., Wierzbiński, J. and Król, G. (2015). Metody ilościowe. In: *Metody badawcze w zarządzaniu humanistycznym*. Warszawa: SEDNO Wydawnictwo Akademickie.
Wieczorkowska-Wierzbińska, G., Król, G. and Wierzbiński, J. (2015). Przeszłość, teraźniejszość i przyszłość edukacji akademickiej. In: *Gospodarka na rozdrożu – XXI wiek*. Warszawa: Wyd. Naukowe Wydziału Zarządzania UW.