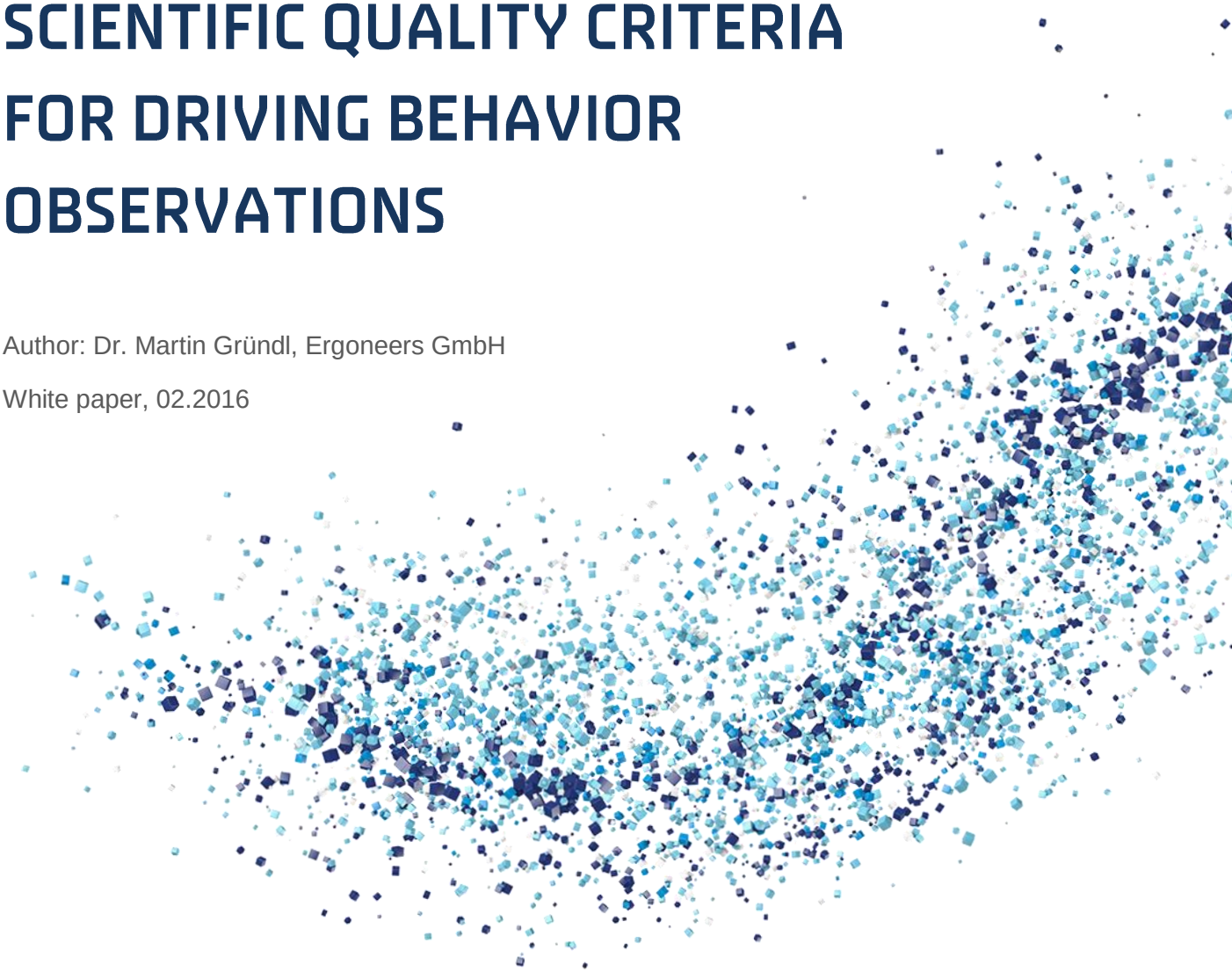
FROM THE DRIVING SIMULATOR TO NATURALISTIC DRIVING STUDIES

PART 1

# SCIENTIFIC QUALITY CRITERIA FOR DRIVING BEHAVIOR OBSERVATIONS

Author: Dr. Martin Gründl, Ergoneers GmbH

White paper, 02.2016



ERGONEERS
FROM SCIENCE TO INNOVATION

# ABSTRACT

In recent years, there has been growing interest in everyday-mirroring driving behavior studies that are not too heavily influenced by experimental conditions. As a result, more and more *“Naturalistic Driving Studies” (NDS)* are being carried out which are designed to capture the driver’s behavior in a natural driving environment. There are also more controlled experimental real-vehicle studies (= *experimental Field Operational Tests, eFOT*), the purpose of which is generally to investigate the effects of driver assistance or driver information systems. The most common method of driving behavior observation, however, remains experiments conducted in the driving simulator.

However no matter which method one might use to observe driving behavior, ultimately the results must always be able to withstand the same scientific quality criteria - namely *objectivity, reliability, internal validity, content validity, external validity* and *replicability*. If these quality standards are not adequately met, a study’s informative merit is restricted - and in extreme cases even zero. In this part 1 of the white paper on the methods of driving behavior observation, the most important quality standards for the scientific investigation will therefore be described and illustrated with examples from the practical field of driving behavior observation.

# CONTENTS

# 1. SUMMARY OF SCIENTIFIC QUALITY CRITERIA

Before new assistance and information systems make their way into cars, automotive manufacturers have to prove through studies as part of a self-imposed obligation (Alliance of Automobile Manufacturers, 2006) that these new systems are safe. This primarily means that they also work correctly in rare, critical situations, that the driver can keep them under control at all times and that he is not distracted from the task of driving too much by using them. To carry out such investigations, there are various methods that differ chiefly in terms of their technical complexity, their closeness to reality and the extent of control over their methodology (for more on this, see **Part 2**).

A trend has recently emerged among these evaluation studies: Standardized laboratory experiments with a controlled study design are being eschewed in favor of gathering driving data in real road traffic in conditions that are as close to reality as possible (Lietz et al., 2001). As a result, more and more *Naturalistic Driving Studies (NDS)* are being carried out which are designed to capture the driver's behavior in a natural driving environment. There are also numerous controlled experimental real-vehicle studies (= *experimental Field Operational Tests, eFOT*), the focus of which is generally to investigate the effects of one or more driver assistance or driver information systems. Here too, the driving behavior is monitored in a natural driving environment, and differences between driving with and without the assistance system (or information system) are detailed. Despite these trends, however, most evaluation studies on driver assistance and information systems continue to be carried out in the driving simulator under laboratory conditions.

No matter which method is used to observe driving behavior, ultimately the results must always be able to withstand the same scientific quality criteria - namely

- Objectivity
- Reliability
- Internal validity
- Content validity
- External validity
- Replicability

If these quality standards are not adequately met, a study's informative merit is restricted - and in extreme cases even zero. In this part 1 of the white paper on the methods of driving behavior observation, the most important quality standards for a scientific investigation will therefore be described briefly and illustrated with examples from the practical field of driving behavior observation. **Part 2** will set out the various methods and their pros and cons explained using the quality criteria described here.

# 2. OBJECTIVITY

A method can be said to have *objectivity* if the results achieved with this method do not depend on *who* is using the method (Klein et al., 2012). If there are two different investigators, then it must be ensured that both behave equally towards the test subjects and give identical instructions, for example. Otherwise the manner in which one investigator explains the task may indirectly give some test subjects useful tips but not others. To ensure as much *implementational objectivity* as possible, it is therefore recommended that instructions be given in writing. During a real journey, for example, the route to be taken can be announced by a sat-nav system, so that the content and timing of the instructions are always exactly the same for all test subjects.

During the analysis of the recorded data, *evaluation objectivity* must also be ensured. If the analysis looks at whether the driver has made a mistake, the result must not depend on which analyst is making the assessment. Whether the driver has braked hard enough in front of a junction at which cars to the right take priority, whether he glanced over his shoulder or not, must be independent of the assessor. The rules for coding the observed behavior must therefore also be correspondingly precise in order to safeguard *objectivity.*

Another example in which *evaluation objectivity* must be examined critically is the measurement of the driver's eye behavior *without* an eye-tracking system. In this instance, only the driver's face is filmed with a camera in the car. Behind this, an assessor (= rater) must analyze the video and evaluate and code where the driver was looking at what point in time (e.g. at the road, at a system display, at the speedometer, at the rear-view mirror, at the wing mirror). Different raters often arrive at very different results. The extent of the agreement between two raters can easily be calculated (= inter-rater reliability). With glances directed at certain areas in the car (= *Areas Of Interest, AOIs*), the category data are convergent. As a statistical measure, *Cohen's Kappa* (Cohen, 1960; Landis & Koch, 1977; Bortz, Lienert & Boehnke, 1990) is used which translates rater agreement into a number between 0 and 1. Glances towards areas of interest that were merely measured with a simple video (i.e. without an eye-tracking device) are a special case of video-based behavioral coding. Unlike when an eye-tracking device is used, there is always a degree of skepticism over just how objective such results really are, and whether the results do not actually depend very much on the rater. By calculating rater agreement, such doubts can be dispelled. However there are not many publications on driver behavior observation that include these statistical parameters for the *objectivity* of the measuring method used.

# 3. RELIABILITY

*Reliability* defines the degree of measurement accuracy (= *precision*) of a measuring method (Bortz & Döring, 2006). Every measurement, however, is also associated with a measurement error. The smaller this measurement error, the greater the *reliability*. The great thing about *reliability* is that - unlike *internal* or *external validity* - it can be quantified. This is because when the same thing is measured twice, the same result should also be reached twice. So if a group of test subjects travels through an identical traffic scenario twice under identical prior conditions, then the driving parameters measured must also be about the same in both cases. If they are very different, something is wrong. Just how well the data match in these two conditions can be calculated using statistical methods, depending on the type of data involved. Correlation is an example of such a method. Although the reliability of a measuring method can be calculated relatively easily - key terms here include retest reliability, parallel test reliability, and split-half reliability - such calculations are carried out far too rarely in driving behavior observations.

# 4. INTERNAL VALIDITY

*Validity* generally means how valid something is. It specifies whether a measuring method is measuring what it is supposed to measure. *Internal validity* relates to the question of how well an investigation is able to control significant disruptive variables. An investigation is therefore *internally valid* if its results can be interpreted in only one way (Campbell & Stanley, 1963; Cook & Campbell, 1979).

During driving behavior measurements, for example, significant disruptive variables include the behavior of other road users, the influence of weather, and visibility conditions. If you want to measure, for example, how a certain assistance system impacts on the safe distance maintained from the car in front and a difference is discovered under the two study conditions “with system” and “without system”, then the differences should also be solely attributable to the test conditions. If however under one set of conditions a shorter safe distance is actually caused by heavier traffic, then the measurement is *internally invalid*. Problems with low *internal validity* due to numerous disruptive variables are typical, especially for studies involving real road traffic. In the driving simulator, on the other hand, it is possible to control third-party traffic and environmental conditions and keep them constant.

Figure 1: Slipping of the eye tracking glasses during a long journey can make the eye measurements for the entire journey unusable, since the eye movement video no longer displays the actual directions of the driver’s gaze. An eye-tracking system with the options of recalibration (such as Dikablis from Ergoneers) can correct this error, however, and therefore increases the internal validity of measured visual data.

Even when an eye-tracking device is used for eye behavior measurements in the car, *internal validity* remains an issue. Does the device also measure where a person is looking, or are the results partially invalidated as a result of disruptive influences? One typical error, for example, is the eye tracking glasses slipping down slightly during a long journey. This causes the initial calibration to no longer be valid and in the eyesight video the test subject’s glances are displayed incorrectly. If an eye-tracking device is used which offers

a recalibration option during visual data analysis, this error can be eliminated. Without this correction option, however, as soon as the glasses start to slip, all eye movements measured are incorrect and no longer *internally valid*.

# 5. CONTENT VALIDITY

Content validity (also known as face validity) is given when whatever you actually want to measure is exhaustively captured by the measuring method in its most important aspects (Bortz & Döring, 2006; Rosnow & Rosenthal, 2007).

Example: A lane departure warning assistant is an assistance system designed to prevent accidents caused by inadvertent drifting from the motorway lane. Shortly before a drive would cross the lane markings (measured as *Time to Line Crossing, TLC*), the system warns the driver or intervenes with a short, gentle movement of the steering wheel. Thanks to an early warning, i.e. the setting of a long *TLC*, a driver can be prompted by the system to make small steering corrections early and therefore drive as close to the middle of the lane as possible. If a driver is driving with this assistance system, he will quickly learn to drive in such a way that he has as few warnings or steering interventions as possible. He therefore remains in the middle of his lane. Without an assistance system, however, he continues to drive as normal, occasionally veering towards the lane markings (i.e. the critical *TLC* is not met). If the two test conditions are now compared in statistical terms, it becomes clear that drivers with the lane departure warning assistant drive more often in the middle of their lane and less often fail to meet the critical *Time to Line Crossing* than drivers without the system. The conclusion that they are also driving more safely and are at lower risk of leaving the road as a result, however, is not necessarily true. This is because failing to meet a TLC of 1.4 seconds, for example, is by no means the same as a near-miss departure from the road, and nothing like near-perfect driving in the middle of the road.

The operationalization in this instance therefore lacks *content validity*. We want to measure the risk of an accident caused by leaving the road, but in fact we are measuring something different, i.e. how exactly and in an exemplary manner a driver can or wants to drive in the middle of a lane. The *content validity* would be greater if the only results counted in both test conditions were those in which the driver *actually* crosses the lane marking. This would be much more informative for an estimate of such a system's accident avoidance potential. However this generally does not happen, because such events are so rare in normal car driving situations that they are very difficult to evaluate statistically. This is because these types of driving studies would consume a lot of time, test subjects and therefore money.

# 6. EXTERNAL VALIDITY

A study has *external validity* if its results can be generalized beyond the particular conditions of the study situation. In this context, a distinction is made between two forms of generalization - namely generalization about *people* and *situations* (Hager & Westermann, 1983; Cook & Shadish, 1994).

If the results are to be generalized beyond the *people* specifically investigated in the study, the random sample of test subjects must be representative. If, due to their more ready availability, only students (= young learner drivers) or only vehicle engineers from an automotive manufacturer (= technology experts) are studied, for example, the results cannot reasonably be transferred to other populations, such as older drivers (= typical buyers of new, high-end cars). Generally speaking, the lower the representativity of the random test subject sample studied, the lower its *external validity*.

If results are to be generalized beyond the specific *situation* studied, this requires, for example, that people in a driving simulator actually behave in exactly the same way as in a real car in real road traffic. It also requires them to behave in a highly standardized experiment in a simulator or in a real driving study with an accompanying investigator in the car in exactly the same way as they would if they were driving unobserved in their own private car. These are very sizable assumptions and, aside from *naturalistic driving studies*, they are also rarely found in typical studies on driving behavior observation. Generally speaking, The more unnatural the study conditions, the lower the *external validity*.

When the observation or measurement influences the behavior being observed or measured and therefore the data gathered, this is referred to as *reactivity*. This reduces the *external validity*. *Reactivity* is therefore a negative criterion. In observation studies, this phenomenon has been referred to for almost 100 years as the *Hawthorne effect* (Roethlisberger & Dickson, 1964). This theory states that the participants of a study will change their natural behavior because they know that they are taking part in a study and that they are being observed. They therefore make more effort to perform better and behave in a more socially desirable manner. In driving observation studies, this means that the test subjects drive with greater concentration, make fewer driving mistakes, ignore distracting peripheral activities and drive more considerately than normal. All of this is to the detriment of *external validity*.

Carrying out a methodologically sound study with a high *external validity* is difficult. This is because there is a conflict of objectives when optimizing *internal* and *external validity*. If *external validity* is optimized, this usually has a deleterious effect on *internal validity* (and vice-versa), so that generally a compromise has to be accepted (Bortz & Döring, 2006).

# 7. REPLICABILITY

Replication is the repetition of a study under identical or very similar conditions. If a study into the effect of an assistance system on driving behavior is repeated, for example, and the same or very similar results are obtained as the first attempt, then the conclusions of the study are significantly more well-founded than without the repetition. If however the results cannot be reproduced, despite the same pre-existing conditions, then the study is *not replicable*. This indicates that, during the execution or evaluation of the studies - even during the first run - some type of disruptive variable has invalidated the result, e.g. a lack of *objectivity*, *unreliable* measurements, uncontrolled disruptive variables such as third-party traffic (i.e. a lack of *internal validity*), or even errors in the statistical analysis of the data.

The tenet that scientific study results must be reproducible is a fundamental principle of science. If results of other, independent researcher groups cannot be reproduced, or if these groups arrive at contradictory results, then the long-term prospect for the conclusions of such studies is an unfavorable one.

The replicability of social sciences studies is generally not particularly good, and in recent years especially there has been heated debate between scientists over the causes of this and what can be done about it (Francis, 2012; Pashler & Harris, 2012; Giner-Sorolla, 2012; Ioannidis, 2012). There is no reason to assume that traffic psychology studies involving driving behavior observations are immune to this problem.

Examples: In real driving studies, disruptive influences from uncontrollable third-party traffic are so great, for example, that random influences can have a major effect on the results. And with conventional random test subject samples of 30 to 50 test subjects, it cannot be assumed either that all of the disruptions that impact at random will somehow average themselves out at the end. The results of *naturalistic driving studies* are even harder to reproduce due to the lack of control. During experiments in the driving simulator, the pre-existing conditions are certainly better, however here too, the results are repeatedly very much influenced by the driving simulator and specific technical settings used. In this case, the replication of a result by an independent group of researchers is made more difficult by the fact that the methods section of publications fails to describe the driving simulator used with all of its settings in adequate enough detail to allow the study to be reproduced under the same conditions.

True replication studies are rare among driving behavior observations. At best, the questions or systems being investigated are the same, but not the entire study context (Makel, Plucker & Hegarty, 2012). There is also a lack of incentive to repeat the studies of other researchers in an identical manner, as well as to critically examine your own findings (Koole & Lakens, 2012). In the automotive sector especially, technology is also advancing rapidly. And even if one researcher were to have a serious interest in replicating a known result, e.g. on the use of a specific driver assistance system, then the automotive industry that usually finances such studies does not. The automotive industry much prefers to have the latest stage of its advanced systems evaluated. The study conditions are no longer the same, thereby precluding true replication.

# 8. SUMMARY

The criteria described are the quality standards for all types of scientific investigations and do not just apply to studies of driving behavior. All criteria are important. As a result,

no single criterion or groups of criteria should be neglected. In practical situations, however, such an ideal is hard to achieve because the optimization of one criterion often leads to a reduction of another. Maximizing *internal* validity, for example, commonly leads to a reduction in *external* validity and vice-versa.

The various methods used in driving behavior observation - from *naturalistic driving studies (NDS)* and experimental real world driving studies to experiments in the driving simulator and simple forms of simulated driving in the laboratory - have all of their characteristic strengths and weaknesses in terms of these quality criteria. **Part 2** of this white paper will describe these methods in more detail and explain their pros and cons in relation to the scientific quality standards mentioned using numerous practical examples. The aim is to provide the reader with a useful decision-making tool for selecting the method most suitable for the purpose of their study.

## Part 2: Pros and cons of various methods used in driving behavior observation

# 9. LITERATUR


Alliance of Automobile Manufacturers (2006). Statement of principles, criteria and verification procedures on driver interactions with advanced in-vehicle information and communication systems, Washington, D.C.: Alliance of Automobile Manufacturers.

Bortz, J. & Döring, N. (2006). Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler. Heidelberg: Springer.

Bortz, J., Lienert, G. A. & Boehnke, K. (1990). Verteilungsfreie Methoden in der Biostatistik. Berlin: Springer.

Campbell, N. & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Cook, T. D. & Campbell, D. T. (1979). Quasi-experimentation. Boston: Houghton Mifflin.

Cook, T. D. & Shadish, W. R. (1994). Social experiments: Some developments over the past fifteen years. Annual Review of Psychology, 45, 545-580.

Francis, G. (2012). The psychology of replication and replication in psychology. Perspectives on Psychological Science, 7, 585-594.

Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. Perspectives on Psychological Science, 7, 562-571.

Hager, W. & Westermann, R. (1983). Planung und Auswertung von Experimente. In J. Bredenkamp & H. Feger (Hrsg.), Hypothesenprüfung. Göttingen: Hogrefe.

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

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

Landis, J. R. & Koch G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33, 159-174.

Lietz, H., Petzoldt, T., Henning, M., Haupt, J., Wanielik, G., Krems, J., Mosebach, H., Schomerus, J., Baumann, M. & Noyer, U. (2011). Methodische und technische Aspekte einer Naturalistic Driving Study. FAT-Schriftenreihe 229, Forschungsvereinigung Automobiltechnik (FAT).

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

Pashler, H. & Harris, Ch. R. (2012). Is the replicability crisis overblown? Three arguments examined. Perspectives on Psychological Science, 7, 531-536.

Rosnow, R. L & Rosenthal, R. (2007). Beginning behavioral research. A conceptual primer. Upper Saddle River, NJ: Pearson Education.

Roethlisberger, F. J. & Dickson, W. J. (1964). Management and the worker. Cambridge, Mass.: Harward University Press.

Ergoneers GmbH was founded in 2005 as a spin-off from the faculty of Ergonomics at the Technical University of Munich. Today the company has a worldwide presence through offices in Germany (HQ near Munich) and USA and through global sales partners; serving the Transportation / Automotive, Market Research & Usability, Science and Sports / Biomechanics application areas.

In addition to development, manufacturing and distribution of measurement & analysis systems for behavioral research and optimization of human-machine-interaction, Ergoneers also offers comprehensive expertise in each phase of your study.

Our product portfolio primarily comprises of the 360-degree solution - D-Lab; an extensive software platform for capturing and analyzing human behavior. With its different software modules you can synchronously measure and analyze eye-tracking, data stream, video, audio, physiology and CAN-Bus data. With the Dikablis Eye-Tracking system, Ergoneers provides the best hardware for professional Eye-Tracking studies in real or virtual environments.

Ergoneers Group
Gewerbering 16
82544 Egling
Germany

T +49.8176.99894-0
F +49.8176.99894-15

Ergoneers of North America, Inc.
111 SW 5th Ave
Suite 3150
Portland, OR 97204
USA

T +1.503.444.3430

info@ergoneers.com
www.ergoneers.com

ERGONEERS
FROM SCIENCE TO INNOVATION