10

How to Fool Yourself With Experiments in Testing Theories in Psychological Research

Werner W. Wittmann and Petra L. Klumb

The Northwestern school, as Glass (1983) coined it, no longer resides in the Department of Psychology at Northwestern University, Evanston, Illinois. Its members are now spread throughout the United States, and its international reputation and recognition is outstanding. Campbell and Stanley (1966), followed by Cook and Campbell (1979), and now by Shadish, Cook, and Campbell (2002), are all the sources that have to be studied, learned, and digested by every student worldwide who wants to do serious research in social sciences. The Northwestern school's influence and impact are still growing. Boruch and colleagues have founded the Campbell Collaboration to promote and foster research synthesis based on randomized experiments and quasi-experiments, especially in the context of education, the field most resistant to experimentation. Cook (2002) analyzed these reasons for resistance. The *American Journal of Evaluation* (2003), in its section "The Historical Record," gives voice to former Northwestern alumni to describe their experiences while being at the university. The number of challengers and critics is also a good indicator of the impact of a school of thought. The Northwestern school has attracted many critics, most importantly, Cronbach (1982; Cronbach et al., 1980), who challenged the preference and emphasis the school has placed on internal validity instead of focusing more on external validity or generalizability of results. Cronbach argued for correlational studies and designs, which may not give the same information about cause-and-effect relationships as the experimental and quasi-experimental designs, but whose predictions are better tailored to real life and give better generalizability. So, the differences between Campbell and Cronbach can be regarded as differences in the emphasis one places on different standards of quality of research designs.

We have been influenced by the debates between the Northwestern school and its critics and have tried to synthesize them into an overall framework.

We thank Guido Makransky and Tobias Bothe for their help with grammar and style.

This allows us, once we know of an evaluation project, to choose between different approaches.

The Five-Data-Box Conceptualization: A Comprehensive Framework for Research and Program Evaluation

For this purpose, we have developed a framework called the *five-data-box conceptualization* (Wittmann, 1985, 2002; Wittmann & Walach, 2002), which refers to five different sources of information one must consider and gather in the process of basic or applied research. Figure 10.1 distinguishes between an evaluation data box (EVA), a criterion box (CR), the experimental treatment box (ETR), the nonexperimental treatment box (NTR), and the predictor box (PR). All data boxes are conceptualized as Cattellian data boxes or covariation charts with their three dimensions: subjects, variables, and situations/time (Cattell, 1988).

The data boxes PR to CR are ordered according to the process of research on a time path. The EVA box on the left contains the stakeholders as subjects.

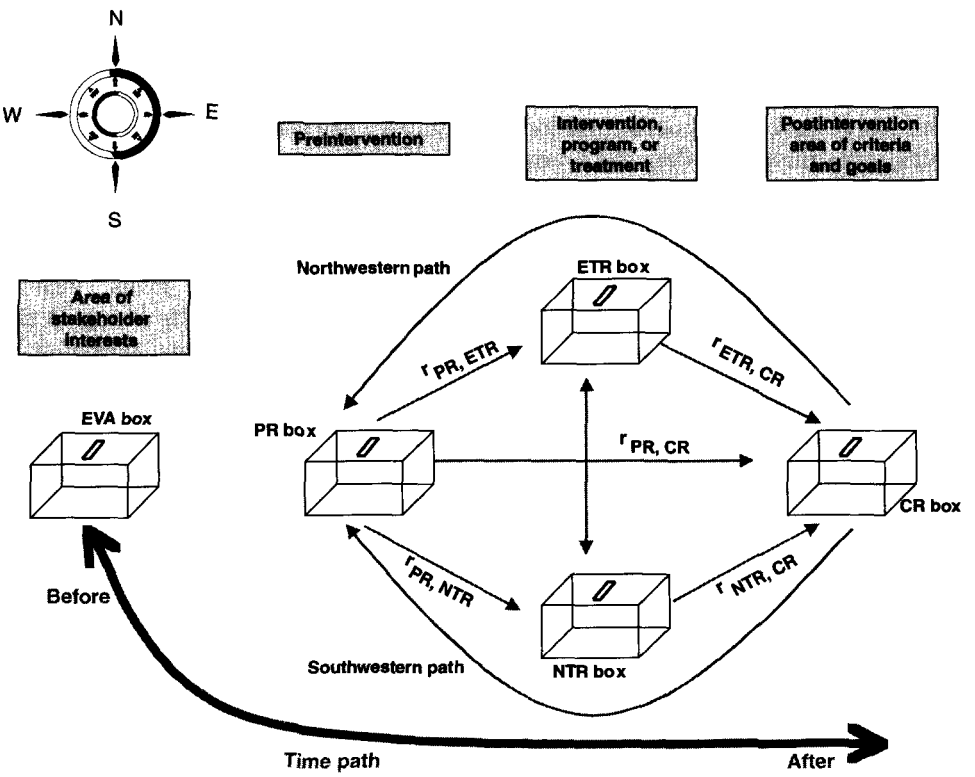


Figure 10.1. The five-data-box conceptualization. EVA box = evaluation data box; CR box = criterion box; ETR box = experimental treatment box; NTR box = nonexperimental treatment box; PR box = predictor box.

Stakeholders are subjects interested in the results—that is, the baseline, the process, the program or intervention, and the impact—of the research. The variables in that box are often fuzzy and vague constructs, which have to be translated into precise measurements by the researcher in program evaluation. In basic research, the subject is the researcher, who is free to choose his or her area of interest. Implicitly, a researcher also must consider one's peers, because difficulties result when there is a lack of interest in the topic that could lead to difficulties in the research being published. The PR box encompasses all variables as baseline data before any intervention. These variables are used for predictions, to control the status before research, and to answer any questions about selection effects regarding the population of interest. The ETR box maps the actively manipulated treatment variables and the members of the randomized experimental and control groups as subjects. In analysis of variance (ANOVA) parlance, these are the independent variables called fixed or random factors and their interactions. The NTR box contains all treatment aspects that could not be randomized—for example, factors mapping nonequivalent comparison groups, such as compliance, dosage, strength, integrity, and fidelity of the intervention. The CR box subsumes all criterion variables, which are used for a summative evaluation of the program or intervention. These variables must map the stakeholder interests and should correspond to what was done as an intervention. Different schools of research and evaluation concentrate on different data boxes and their possible relationships. If we regress the CR box on the ETR box and the PR box, we follow the Northwestern path. If we regress the CR box on the NTR box and the PR box, we follow the Southwestern path. The geographical wind rose at the upper-left corner of Figure 10.1 serves as a guide to reading the data-box conceptualization as a geographic map to facilitate our understanding of the contrast between the Northwestern schools and the Stanford evaluation consortium (with Lee Cronbach as the main spokesman) in what they consider important and feasible in program evaluation.

Suchman (1967), in the first systematic textbook on evaluation research, put the highest priority on the Northwestern school. He considered Campbell and Stanley (1966) as the "bible" of the researcher. Unfortunately, many evaluation studies showed low or zero effects. Rossi (1978) referred to that state of affairs as the iron law of program evaluation. The stately mansion of evaluation research and program evaluation rests on three strong pillars, namely, research design and the related data-analytic tools of assessment methods and decision aids.

Lee Sechrest has contributed to assessment (Sechrest, 1986; Sechrest, Schwartz, Webb, & Campbell, 1999), to debates about quantitative versus qualitative research, and to the problems related to treatment integrity, fidelity, implementation, and strength (Sechrest, West, Phillips, Redner, & Yeaton, 1979; Sechrest & Yeaton, 1981, 1982; Yeaton & Sechrest, 1981). Lack of treatment integrity or failures in implementing a program can easily explain why the program did not show the effects its stakeholders hoped. Boruch and Gomez (1977), in the same sense, proposed a small measurement theory in the field and pointed to the problem of overlap between treatment and its outcome measures as an explanation for low or zero-order effects.

The debates about adequate research designs and its data-analytic strategies have a long history in psychological science. In the 1950s, there was a heated debate between proponents of experimental design and those of representative design. Egon Brunswik (1955), who proposed the representative design, was heavily criticized, especially by Hilgard (1955). Brunswik's data-analytic tool was regression/correlation. It is well known that correlations are only a necessary, but not sufficient, prerequisite for causal explanations. Yet when the time paths are known, we can use regression analysis as path analysis (Wright, 1921) to search for true causal relationships, even in nonexperimental designs, distinguishing between direct and indirect effects and false causal claims as spurious. We can control for selection into treatment effects but still have to face the problem of generalizability and the possible consequences of unmeasured causes. Experiments are traditionally analyzed with Fisher's ANOVA, and many researchers believe that doing an ANOVA brings them all the virtues of a randomized experimental design. Cohen (1968), in his seminal paper, demonstrated that all designs, whether experimental, quasi-experimental, or plain correlational, can be analyzed by the general linear model—that is, multiple regression/correlation. His paper was expanded into a full textbook (Cohen & Cohen, 1975), which has seen its third edition (Cohen, Cohen, West, & Aiken, 2003).

The four boxes on the right-hand side of Figure 10.1 are related with directed arrows mapping the time paths between them. Only the relationship between the ETR box and the NTR box is denoted with a double-headed arrow, indicating the gradual decline from a fully randomized design to a more quasi-experimental and correlational one. It is interesting to note that the title of Cook and Campbell (1979) already was *Quasi-Experimentation*, demonstrating that the Northwestern school was fully aware of the problems associated with doing research in field settings (i.e., real life). Nevertheless, Cronbach (1982) accused the Northwestern school of putting too much emphasis on internal validity and neglecting external validity or generalizability. Cook (1993) and Matt (2003) are Northwesterners most open to Cronbach's challenges, and Shadish et al. (2002), in the latest completely reworked edition of the "research bible," integrate ideas about generalizability and how to better balance conflicts between internal and external validity.

Brunswik Symmetry: A Key Concept for Successful Psychological Research

Looking for reasons that natural sciences like physics, chemistry, and biology have been so successful, we often find references to the experimental methods and good falsifiable theories. It is no wonder that those ambitious enough to change psychology from literature and art to science insisted so much on the experimental approach. However, psychologists have neglected another key concept for success in science, namely, the ubiquitous concepts and principles of symmetry. Zee (1989) described symmetry, and we have learned that the successes in physics of Michael Faraday, Murray Gell-Mann, and Richard Feynman, among many others, would not have occurred without capitalizing

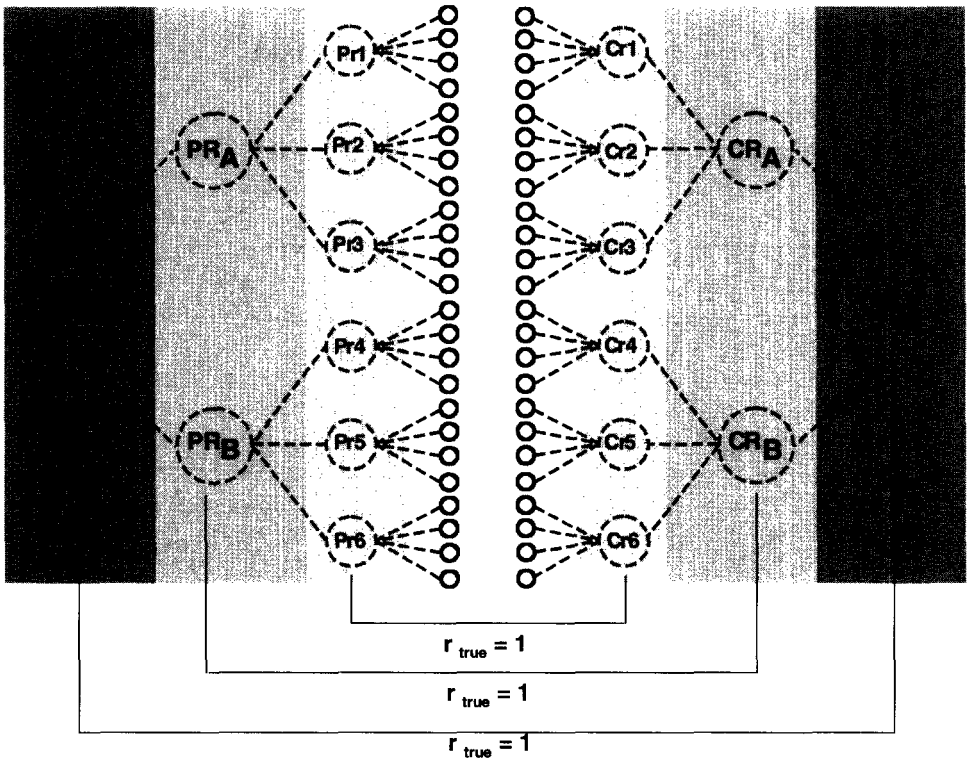


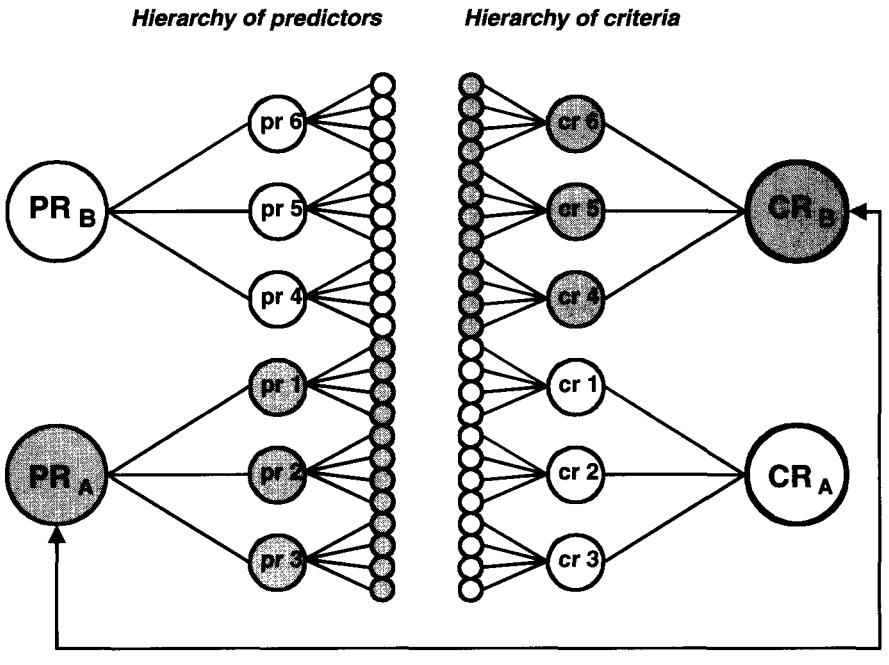
Figure 10.2. The true Brunswik symmetrical latent structure of nature. PR = predictor; CR = criterion.

on symmetry. Brunswik's main conceptual breakthroughs—the representative design and the lens model for human perception and judgment—have not been appreciated by most of his peers, but his ideas have survived with the help of Hammond (1966, 1996; Hammond & Stewart, 2001). We focus on his lens model and use it to look at the relationship between our data boxes. Figure 10.2 visualizes the PR–CR box relationship.

The Gestalt principles immediately force us to consider symmetry principles in amount of aggregation, level of generality, and correspondence between predictor and criterion constructs. Only when these principles hold can we hope to get maximum validity in terms of correlation coefficients or variance accounted for. Variants of violations of symmetry give us hints to how and when our research might fool us. Figure 10.3 distinguishes four variants of asymmetry.

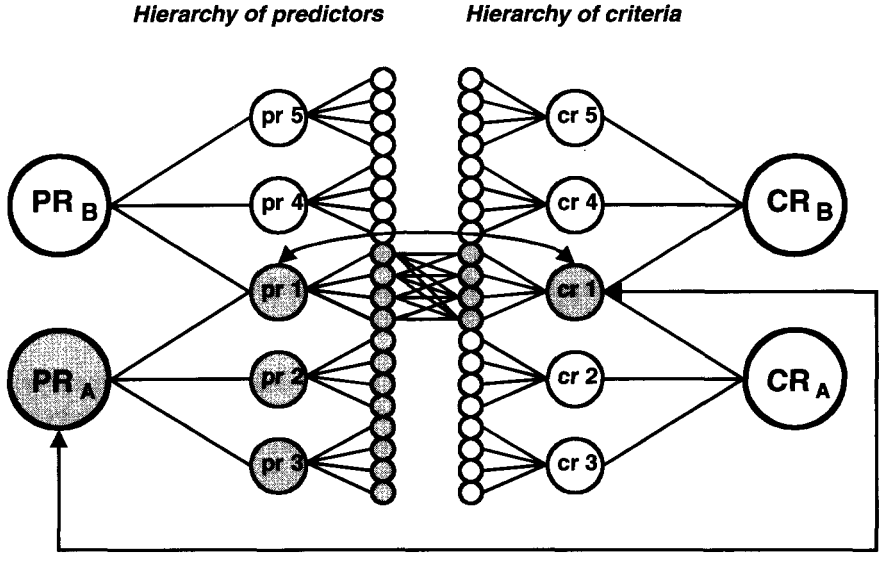
Figure 10.3a shows the case of full asymmetry, which is the case in which nothing works. Predictors and criteria do not overlap; it is the extreme case when what is taught and what is tested do not correspond. The reliability of the predictor and the criterion constructs may be perfect, but we have no

(A)



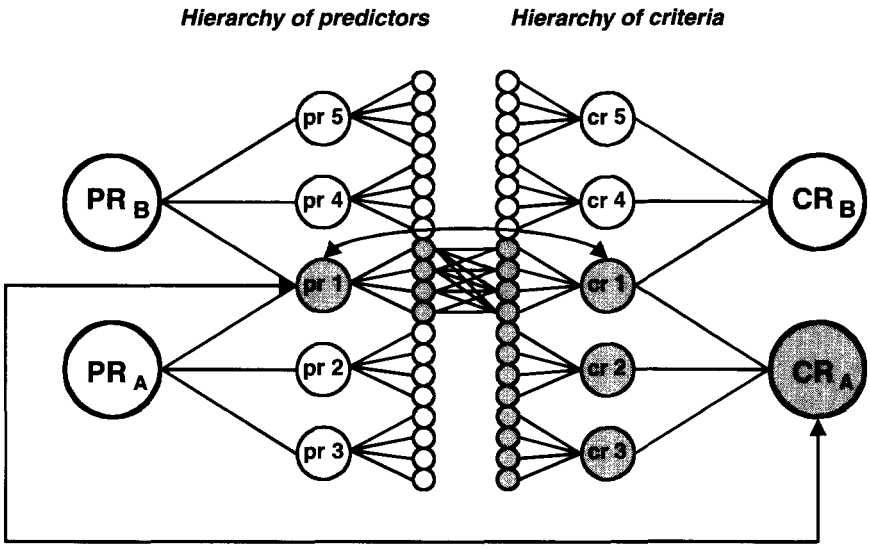
All correlations between predictors and criteria are zero!

(B)



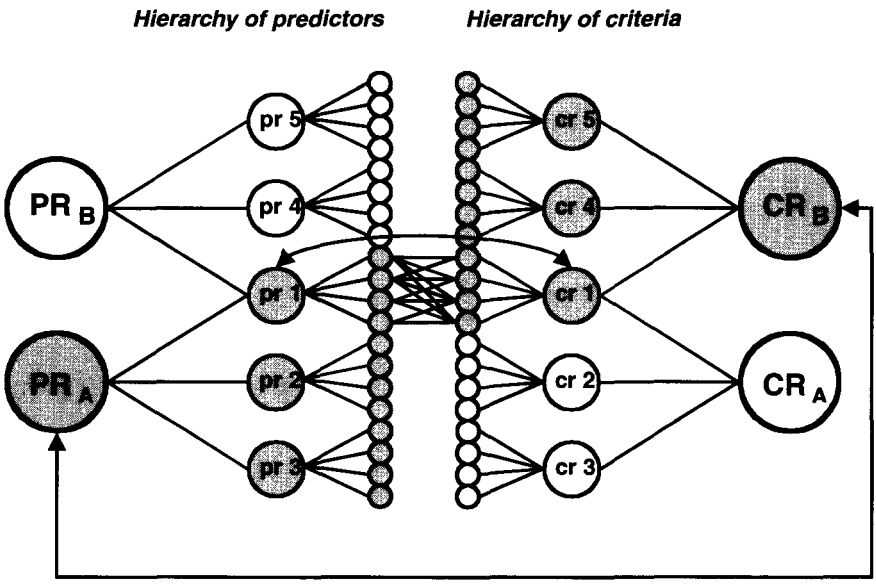
Predictor and a narrower lower level criterion.

(C)



Predictor and a broad higher level criterion.

(D)



Mismatch at the same level of generality!

Figure 10.3. (A) Full asymmetry: the case in which nothing works. (B) Asymmetry because of a broad higher level predictor. (C) Asymmetry because of a narrower lower level predictor. (D) The hybrid case of asymmetry. PR = predictor; CR = criterion.

predictive validity. This case happens by choosing assessments according to their psychometric reliability only and not in terms of their construct relevance, or, as we call them, their construct reliability. Nevertheless, it is an interesting case because, according to Campbell and Fiske (1959), we have perfect discriminant validity. Knowing what something is not is very helpful for falsification in a Popperian sense and serves for construct validation. Figure 10.3b denotes the case in which we have a broad predictor construct and a narrow criterion; they do not correspond in nomothetic span. This case illustrates the problems in the Epstein–Mischel debate about the validity of personality trait dispositions. Epstein (1980, 1983; Epstein & O'Brien, 1985) focused on the importance of aggregation and demonstrated that he could boost on validity, but Mischel (Mischel & Peake, 1982) insisted on the predictability of behavior in the specific situation.

Figure 10.3c illustrates the case of narrow predictor and broader criterion constructs. This case has a sad tradition in psychology. Applying construction principles of homogeneity in assessment via Cronbach alpha or Kuder–Richardson estimates, we drill a smaller and smaller hole into a construct, gaining internal consistency reliability but losing nomothetic span. Many of our assessment tools derived this way later show chronically low validity because they have lost the nomothetic span of criteria we are interested in. Figure 10.3d is the hybrid case, in which we have a mismatch at the same level of generality (i.e., only partial overlap). Validity is different from zero, but is this an indication of convergent or discriminant validity?

This visualization is immediately evident, and it is easy to find examples in which we might have fooled ourselves. We can apply the same principles to the relationship between the treatment boxes and the criterion box. Doing this, we ask how the intervention is operationalized or assessed. Figure 10.4 shows the ETR box.

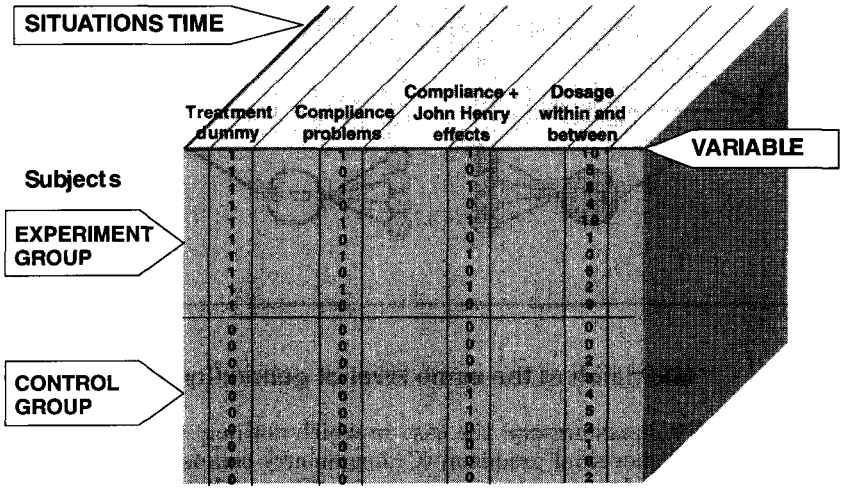


Figure 10.4. A closer look at the experimental treatment box.

Opening that black box, we find, for the randomized experimental control group design, a single dummy variable only, contrasting the experimental group with numbers 1 with the control group with numbers 0. This is a poor and crippled assessment from the stance of a psychometrician when we consider the treatment being a comprehensive intervention or a whole treatment package or program. What about maintaining treatment differences over time? What about dosage differences? What about treatment integrity and fidelity? What about delivering the treatment as intended? It is another irony or paradox that so much is invested in measuring the dependent variables but forgotten for the independent variables in experiments. What insights result if we look at the independent variable in a typical experimental design from a psychometric stance? What is its reliability? Wittmann (1988), in a multivariate reliability theory, proposed a solution and equations, but we have found no application of that concept so far. Reliability is defined as true variance divided by observed variance. True variance is the systematic variance between groups, and the observed total variance is variance between groups plus variance within groups. Looking at the treatment/control dummy (Figure 10.4), we immediately see that the pooled variance within groups is 0. Thus, an experimenter implicitly assumes that the reliability of the independent variable is always 1! But this is wishful thinking, because of compliance, implementation, John Henry effects (compensatory rivalry), and dosage problems, among many others. We can anticipate that there must be variance within groups, but how large is that variance? Good experimental planning asks for manipulation checks. Unfortunately, these manipulation checks test whether there is any difference between the experimental and the control group only. Often, chi-squares are used for that purpose. With a significant chi-square, we know that the manipulation was successful, but we know little about reliability, except that it is different from zero. To find how much an effect size is attenuated, we must compute that coefficient. In some examples discussed later, we found that reliability was chronically low. Lack of power to detect an effect when it is there is the inevitable consequence. According to Cronbach (1957, 1975), this is another consequence of the two disciplines of scientific psychology. He thought more about the conceptual problems, but the two disciplines also had developed their own favorite tools and failed to synthesize them. Cohen's (1968) seminal paper also took a long time until it was finally brought into data analysis. This caused most graduate programs to teach only ANOVA, which in turn caused the next generation of researchers to learn little about multiple regression/correlation, the general linear model, and how it can be used to analyze almost every design. Those who learn both methods risk wasting a great portion of their time.

The principles of symmetry related to Brunswik's lens model cannot be assessed either verbally or visually alone, but require a mathematical numerical equation, thanks to an elegant solution given by Tucker (1964). Here is the original form of that equation.

$$r_{PR,CR} = G_{PR,CR}R_{PR} \cdot R_{CR} + C_{PR,CR}\sqrt{(1 - R^2_{PR})(1 - R^2_{CR})} \qquad (1)$$

The observed predictor–criterion correlation is explained by several parameters. The first part is related to a linear model and the second part to a model

that contains nonlinear aspects and random error. R_{PR} and R_{CR} are linear models of the predictor and criterion, respectively; they have to be computed by regressing a higher level construct on its lower level indicators. $G_{PR,CR}$ is the correlation between these two linear models. The terms $(1 - R^2_{PR})$ and $(1 - R^2_{CR})$ contain variance not accounted for by the linear model; thus, they map all systematic nonlinear variance and error. Parameter $C_{PR,\,CR}$ is the correlation between the nonlinear models of both sides in the sense of orthogonal polynomials, where the linear models already have been partialed. In developing Equation 1, Tucker gave a helping hand to those analyzing problems in human judgment and decision making, but his equation has much more generality, and we consider it as the most important equation psychology has developed thus far. From psychometric theory, we know that no measures are perfectly reliable and correlation coefficients may vary because of selection effects and sampling error, so we simply augmented these concepts into Tucker's lens model equation. Equation 2 shows this augmented equation for the relationship between the ETR box and the CR box, because our focus here is on how we can fool ourselves with experiments.

$$r^{\text{observed}}_{\text{ETR,CR}} = S_l \sqrt{r^{\text{ETR}(l)}_{\text{tt}} \cdot r^{\text{CR}(l)}_{\text{tt}}}\, G^{\text{true}}_{\text{ETR}(l),\text{CR}(l)} \cdot R_{\text{ETR}(l)} \cdot R_{\text{CR}(l)} + S^{\text{n}} \sqrt{r^{\text{ETR}(n)}_{\text{tt}} \cdot r^{\text{CR}(n)}_{\text{tt}}}\, C^{\text{true}}_{\text{ETR}(n),\text{CR}(n)} \cdot R^{\text{ETR}(n)} \cdot R_{\text{CR}(n)} + e \tag{2}$$

The additional parameters are as follows: The terms $r^{\text{ETR(l)}}_{\text{tt}}$ and $r^{\text{CR(l)}}_{\text{tt}}$ are the classical psychometric reliabilities of the linear models of the operationalization of the experimental treatment and the criterion, respectively. The terms $r^{\text{ETR(n)}}_{\text{tt}}$ and $r^{\text{CR(n)}}_{\text{tt}}$ are the psychometric reliabilities of the nonlinear models, and e stands for error. S_l and S_n mean linear and nonlinear models denoting selection effects. Dawes and Corrigan (1974) demonstrated the robust beauty of linear models in psychology and the social sciences, so we simplify Equation 2 by dropping the nonlinear term. Parameter S is equal to 1 only when the sample standard deviation is equal to the population standard deviation; when SD_{sample} is smaller than SD_{pop}, S is smaller than 1; and when SD_{sample} is larger than SD_{pop}, S turns out to be larger than 1. S is only a placeholder to denote the selection problems that have been known since Thorndike (1949). Hunter and Schmidt (1990) gave the following equation:

$$r_{\text{sample}} = u\, r_{\text{pop}} \sqrt{(u^2 - 1)\, r^2_{\text{pop}} + 1}, \text{ where} \tag{3}$$

$$u = SD_{\text{sample}} / SD_{\text{pop}}. \tag{4}$$

To demonstrate how large S gets under selection, we have constructed a nomogram for a rough calculation of these effects (Figure 10.5).

The abscissa shows u and the ordinate r_{sample}; for r_{pop}, we have chosen small (.10), medium (.30), and large (.50) effect sizes (Cohen, 1992). Restriction of range occurs when $u < 1$ and enhancement of range when $u > 1$. For small effect sizes in the population, there is a linear relationship: The larger the effect size, the more nonlinear the effect of u is. When the standard deviation in the sample is only half of the standard deviation in the population (i.e., u = .50), with a large effect size, we get only a sample effect size = .28, and S would

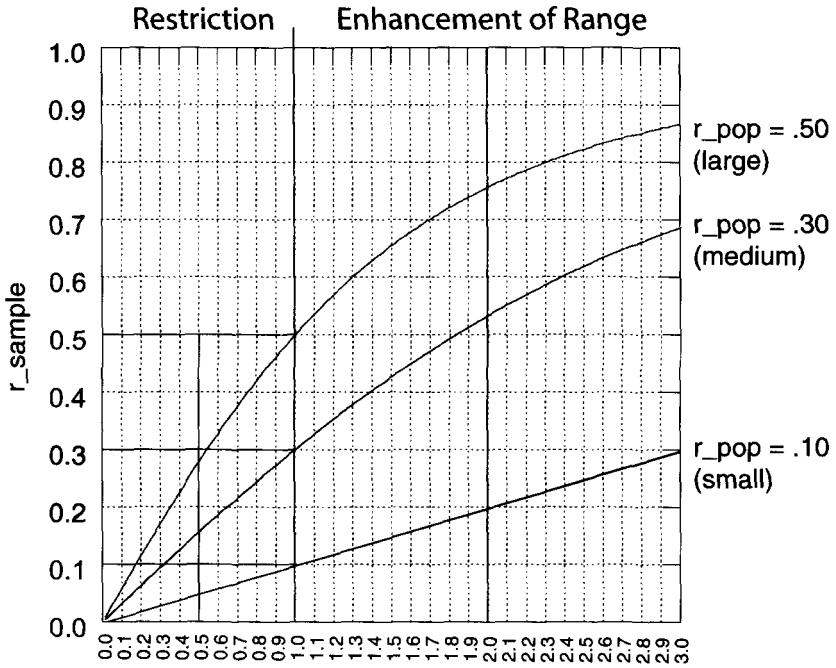


Figure 10.5. Nomogram for selection effects: parameter S.

be .28/.50 = .56. If u = 2.0, then the sample effect size is roughly .76. S is then .76/.50 = 1.52; it tells us how much we overestimate the effect in the population. To underscore the importance of the modified Tucker lens-model equation, it is shown again in its linear parts as Figure 10.6.

The true effect size in the population is surrounded by parameters that either attenuate or augment it. There are six dangers of underestimating a true effect and only two dangers of overestimating it. Therefore, the odds of underestimation are higher than those of overestimation! This is an important lesson and gives an idea about how much psychology has fooled itself in regard to its research results. The observed effect sizes are used as a decision aid to evaluate the impact and worth of psychological strategies and interventions. Fortunately, we now have meta-analysis for these summative evaluation purposes. Glass (1983), Hunter and Schmidt (1990), and Rosenthal (Rosenthal, Rosnow, & Rubin, 2000), among many others, have contributed to popularizing meta-analysis. Glass synthesized experiments in psychotherapy, Hunter and Schmidt started synthesizing validity coefficients in personnel selection research and termed their approach "validity generalization," and Rosenthal synthesized the p values from significance testing. All these approaches are now under a common framework. (See Rosenthal et al., 2000.) The d and r families of effect sizes can easily be transformed into one another. The effect size r can be transformed into Cohen's d as

$$d = r / \text{sqrt}\,(pq(1 - r^2)), \tag{5}$$

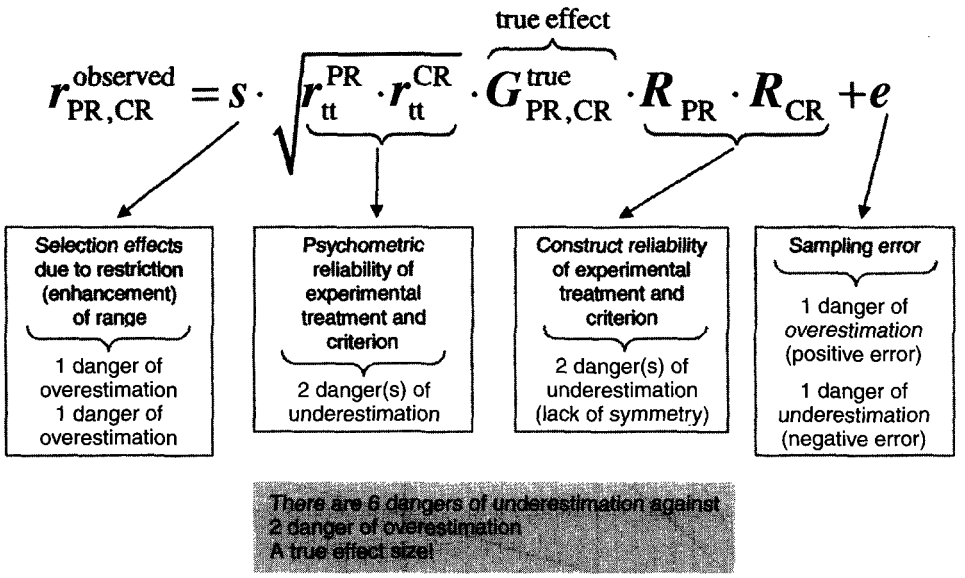


Figure 10.6. The Brunswik lens equation for relating experimental treatment (ETR) to criteria (CR).

where p and q are the proportion of participants in the experimental and control group, respectively. For $p = q = .50$, where we have the same number of participants randomized to both groups, we get the simplification $d = 2r/\text{sqrt}(1 - r^2)$. Inserting Equation 3 into Equation 5, we would learn how much d is attenuated or augmented by the research artifacts discussed earlier.

For the experimental approach, we must reflect on what the possible distribution of the independent variable is. Is it normally distributed, rectangular, or something else? Causes do not have a distribution; they differ only in dosage level or strength. In asking what the right dosage is, we know that dosages too high are often lethal or could be a waste of effort. Lipsey (1990, 1993) discussed the independent variable and the role of theory. He distinguished five different types of dose and response relationships, which differ by the onset process of an effect as a function of dosage. The first is a step function mapping a sharp and maximal onset, the second and third nonlinear functions mapping effects for strong or weak doses, the fourth and fifth U-shaped and inverted-U functions. These theoretical considerations are important in realizing the MAXMINCON principles recommended by Kerlinger (1973), which state that one should maximize (MAX) the effect between groups but minimize (MIN) the variance within and control (CON) for unwanted systematic variance. The experimental and control group must differ in the dosage level, and the split by which we map our treatment dummy must correspond to that level at which we assume that an onset of the response occurs. For such unitary causes, we need a lot of theoretical knowledge about where to make the split. In most program evaluation, whole treatment packages are the interventions, and we can assume that several causes should be at work. Whatever the dose–response

functions of the unitary causes are, the composite-cause distributions are probably normal again, so few people will receive a low and few a high composite dose, and we again can hope to profit from the robust beauty of a linear model assuming a linear relationship between response (most often also a normally distributed composite) and composite dosage. Now the question of where to make the split in complying with the MAX principle brings us back to the problems of enhancement of range mentioned earlier. The popular strategy of using extreme groups from both tails of the composite cause brings more power into the design but gives no answer to whether we can generalize such an effect. Nevertheless, once we know parameter S, we can correct the effect we find in such designs after we implement the program to the full population and can guess whether such an effect would be worth the investment. Restriction-of-range problems have their mirror in thinking about how much the psychology students used in our experiments represent the full population. Cohen (1983) warned us about the cost of dichotomization of a normally distributed variable. Assuming a normally distributed composite, he demonstrated a proportional loss of .80 once we make the split at the median; splits farther away from the median result in a still more dramatic reduction of effect size and the inevitable loss of power.

The main point of all these considerations is that psychology is under the permanent threat of underestimating the effects of all types of its interventions and strategies it has developed thus far. Cohen was much depressed finding that the power of the research design to detect medium effect sizes had declined from .48 (Cohen, 1962, 1977) to .25 when Sedlmeier and Gigerenzer (1989) reported their second look at research results.

Meta-Analysis and the Brunswik Lens Model Equation

Hunter and Schmidt (1990) used the parameters of Figure 10.6 to investigate how far the variability in the parameters around the true effect can explain the variability of observed effect sizes. They proposed the 75% rule, meaning that when 75% of the variance of observed effect sizes can be explained by these artifacts, the overall effect can be generalized and there is no need for looking additionally at moderators that can explain the true effect variability. They used this rule mainly for personnel selection research, which is represented by the relationship between the PR and the CR boxes in the five-data-box conceptualization. Their conclusion was that in this area the 75% rule is given, and so far, one can generalize the validity coefficients of the tests used. Consequently, there is no need to validate them anew in each selection situation! Smith, Glass, and Miller (1980) also investigated whether selected aspects of research quality are correlated with effect sizes resulting from the experimental designs used in psychotherapy research. They found no substantial correlations. Wittmann and Matt (1986) looked at German-speaking psychotherapy research only and used a more extended rating scheme of quality according to internal, statistical conclusion, external, and construct validity (Cook & Campbell, 1979); they also distinguished the construct validity of causes and effects and differences in external validity (e.g., do the intentions to generalize

correspond to the design used?). This Northwestern rating scheme unraveled substantial correlations with effect sizes. When only the variables used by Smith et al. (1980) were analyzed, there were no substantial correlations, thus replicating their results even in German-speaking psychotherapy research, but this also meant that the extended rating of quality made a difference (Wittmann, 1985, 1987b). Behavioral interventions had higher effect sizes compared with psychodynamic ones. The main reason for that was the use of assessment instruments in the CR box. The former better tailored these instruments to what is treated and what is tested, more behavioral checklists and instruments thought to be sensitive to change in the first place, whereas the latter more often used broad dispositional personality scales based on trait theory and trimmed to stability aspects of behavior. Therefore, the psychodynamics fell more than others did into the asymmetry trap visualized in Figure 10.3c. A lead indicator was whether the design was designed a priori as a follow-up study, taking a larger slice of the time–situation coordinate of the CR box. Those who did design it that way had a better hypothesis about the stability of effects and their generalizability over time, used multimethod and multivariate assessments, and focused more on specific aspects of personality and specific subgroups. One can speculate that when a follow-up design with extended postmeasures over time is used, the researchers would already have accumulated more knowledge about causal effects, making them confident that the intervention works. Otherwise they would not have invested the extra resources these designs require.

With regard to the importance of design validity, we found significant correlations for all four Northwestern standards, but the construct and external validity were relatively more important than internal and statistical conclusion validity. This sheds an interesting spotlight on Cronbach's stance discussed earlier.

To test the Brunswik symmetry principles, we built an index mapping symmetry between the causes and effects in terms of external and construct validity, with low scores indicating high symmetry and high scores indicating high asymmetry. Figure. 10.7 shows effect-size box plots as a function of asymmetry, and the overall distribution bolsters our hypothesis.

Secondary Analysis of Three Selected Research Studies

Encouraged by the promises of the Brunswik symmetry framework, we took a second look at three different single-research studies. The first is a longitudinal study of Fahrenberg, Myrtek, Kulick, and Frommelt (1977) sampling behavioral observations over 8 weeks, which we used as an attempt to validate Eysenck's personality theory (Wittmann, 1987a). The second is a program evaluation study by Lösel, Köferl, and Weber (1987) about the training effects of prison officers (Lösel & Wittmann, 1989). The third is a comprehensive quasi-experimental study by Klumb (1995) to test the validity of a questionnaire related to Donald Broadbent's memory-based theory of cognitive failures and lapses.

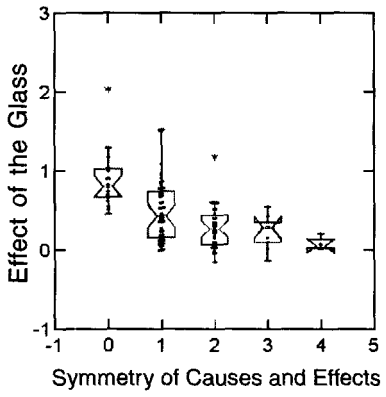


Figure 10.7. German psychotherapy effects as a function of symmetry. 107 effects from Wittman and Matt (1986) and the extension by Spinner (1991). Low scores represent high symmetry.

The Promise of Longitudinal Designs for Personality Traits (Fahrenberg et al., 1977)

Fahrenberg's lab at the University of Freiburg, Germany, is most well known for its focus on psychophysiology. Fahrenberg also developed the most-used German-speaking personality inventory, the *Freiburger Persönlichkeits-inventar* (FPI). The FPI (Fahrenberg, Hampel, & Selg, 2001), among other dimensions, also measures Eysenck's extraversion and emotional lability (neuroticism) factors. In the study, we assessed 20 students over 8 weeks. At the beginning, the students took the FPI, and over the 2-month period they kept daily diaries with many behavioral observations and self-ratings. Two times per week, they visited the lab, where they took psychophysiological assessments and were rated by the researchers. In the secondary analysis, we scanned Eysenck's research and literature about what he claimed to be indicators of extraversion and neuroticism. We found eight indicators for extraversion and seven indicators for emotional lability in Fahrenberg et al.'s study. From a theoretical stance, we assumed that traits are dispositional constructs. A disposition is a tendency to act in a specific situation (here, a day) in the direction of the dispositional construct. We do not expect that the postulated behavior will show up consistently in each situation, but in the long run those high on the trait should show the behavior or feeling more often than those with low scores on the construct. This postulates higher Brunswik symmetry of traits with aggregated criteria over time. Brunswik symmetry in this case is nothing more than the principle of correspondence in target, context, action, and time proposed by Fishbein and Ajzen (1975) in attitude research. They proposed to distinguish among single acts, repeated single acts, and multiple acts in a relatively specific situation or time frame and repeated multiple-act criteria (RMAC), which aggregate functionally equivalent behaviors and feelings

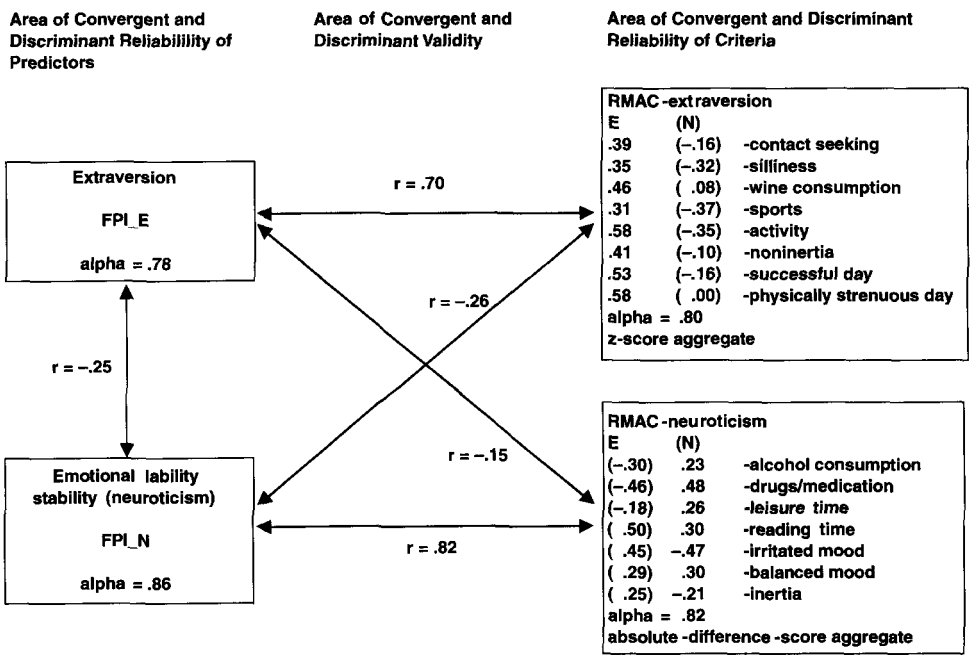


Figure 10.8. Testing Eysenck's Extraversion–Neuroticism theory in the Brunswik symmetry framework. Time series data of 20 students assessed over 8 weeks (from Fahrenberg et al., 1977). FPI = the German-speaking personality inventory *Freiburger Persönlichkeitsinventar*. RMAC = repeatedly measured multiple-act criteria.

(RMAC) over many situations or periods. For the extraversion RMAC, we could aggregate over 60 days. The RMAC for emotional lability was constructed by means of absolute difference scores. For these indicators, we first computed mean level for each half-week and then an absolute difference score per week, which then was aggregated over all 8 weeks. The reason was dictated by the meaning of the construct: Lability should show up as variability, and the absolute difference scores are an attempt to assess the ups and downs over a longer time. Figure 10.8 shows the results.

Applying Campbell and Fiske's (1959) principles of convergent and discriminant validity yields impressive results. Eysenck's theory postulates extraversion and neuroticism to be independent. The low correlation in this sample is not significant; in addition, the discriminant validity coefficients are insignificant and the convergent validity coefficients are impressively high, much higher than what Mischel (1968) had coined as a personality coefficient. Almost perfect Brunswik symmetry would result with the reliability estimates for correction for attenuation. Although we are aware of the limitations of a sample size of 20 and the dangers of generalization to the whole populations of either students

or all persons, the generalizability over time is impressive. The results also hint at a possible solution of the Epstein–Mischel debate. Personality traits might be very good predictors for aggregated multiple-act criteria but not so good for a specific single act. However, we still have to wait for answers to what brings the same amount of predictive validity for situation-specific behavior; that is, what are the decisive situational characteristics, despite the massive restructuring of the majority of psychological departments in the world favoring social psychology? The study had neither an ETR nor an NTR box, but we can nevertheless speculate what must be done once we think about changing these traits. Because of the multifaceted criteria and the predictive success, we can assume that Eysenck's factors are multifaceted as well. So, to change them, we need a corresponding symmetrical intervention, which can only be a multifaceted treatment package. We saw that variability in the use of alcohol, drugs, and medication plays a role. It was not the mean level in these parameters but their ups and downs, so what triggers their onset? How should we deal with relapse prevention? How can we stabilize mood variability? Should we use medication or cognitive–behavioral interventions? How can we deal with the variability in leisure time? What are the right treatments to better balance social activity with retreat? An experienced clinician should get many hints on how to package a comprehensive composite treatment to change these traits, given that subjects regard them as a problem.

Training Prison Officers With Psychological Interventions (Lösel et al., 1987)

Prison officers are the people who have the highest amount of contact with prisoners. Therefore, training and supplying them with helpful skills should be a promising strategy to empower them as change agents. Behavior therapy and Rogerian types of intervention have a lot to offer for changing behavior, emotions, feeling, and interpersonal skills. Four trainers with a behavioral therapy background and four trainers with a Rogerian background were used. They were partially randomized and matched to train and educate 11 or 12 prison officers in each group. The groups were compared with each other and with an untrained control group. The program-centered training (PCT) groups followed the tradition of behavioral learning theory, whereas the group-centered training (GCT) groups followed the tradition of T-group laboratories. As criterion measures, theory-derived outcome variables were chosen to map effects, which can be best expected on the basis of what each intervention trains. Attitudes toward using psychological knowledge in prison and reactions in specific test situations emphasizing behavioral competencies and communicative sensitivity were used as criterion variables. The first two are closer related to what was taught in the PCT groups, and communicative sensitivity is closer to what was taught in the GCT groups. The training took 1 week, all training sessions were videotaped, and the posttests were given 5 month after training. Data analysis showed no significant differences between the PCT and GCT groups on the first criterion. The effect size was $r = .11$, $t(92) = 1.08$. In

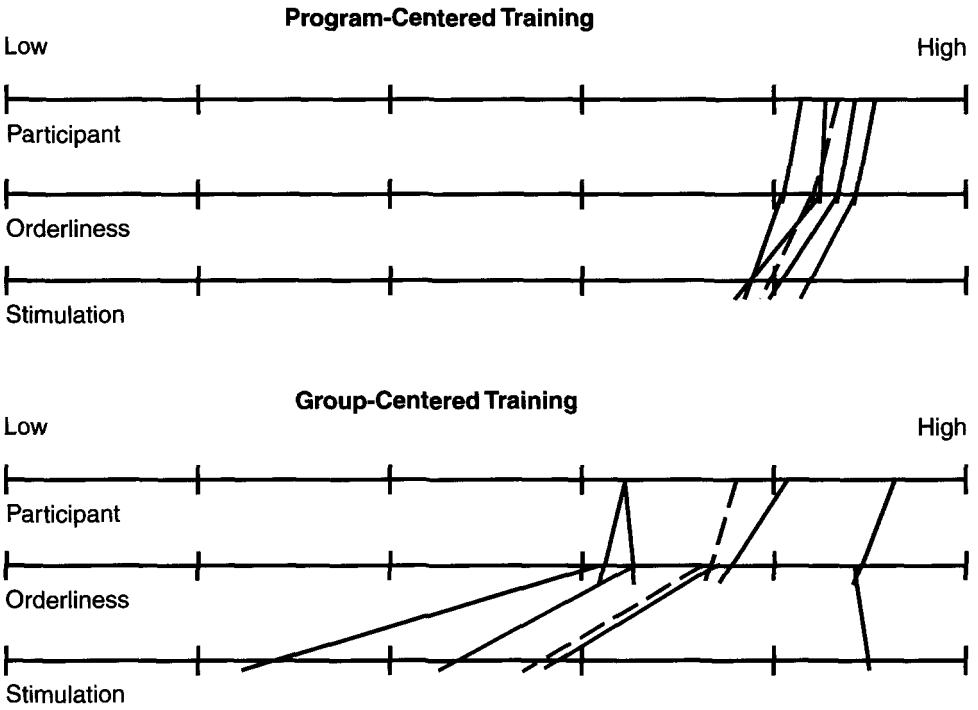


Figure 10.9. Behavior of group trainers as perceived in single courses (plain lines) and on the average (dashed lines).

the second criterion, the effect size was $r = .06$, $t(91) = 0.53$, *ns*, and neither was significantly different from the control group. For the third one, most relevant for the GCT group, there was a significant difference from the control group but no significant difference from the PCT group. Effect size here was again $r = .11$, $t(91) = 1.09$. The summative evaluation would have ended as another example of no-difference research or an additional study to bolster Rossi's (1978) iron law of program evaluation had we not taken a closer look at treatment integrity. All videotaped training sessions were process evaluated by time-sample analysis. As indicators of integrity and intensity, three dimensions assessing trainer behavior from the video time samples were rated and aggregated over all time samples according to participant orientation, orderliness, and stimulation, following Ryans (1960). The results for the eight courses are shown in Figure 10.9.

As can be seen, the PCT group is rated more homogeneous and with higher average intensity on all three dimensions. Within the GCT group, one course is an outlier and seems to be a most intensive PCT course despite this psychologist being hired as a GCT trainer. Additional information about amount of speech and emotional qualities also confirmed that this trainer was closer to PCT than to GCT. Applying our equation for treatment reliability, we found coefficients of .38 for participant orientation, .48 for orderliness, .33 for stimulation, and

.38 for the total scores over all three dimensions. Obviously, realizing Kerlinger's (1973) MAX principle was not successfully established; treatment homogeneity within groups was lacking. As Figure 10.9 hints, the main reason was the GCT trainer, who behaved as a PCT trainer. Regrouping his sessions to PCT and recalculating the treatment reliability brought coefficients of .80 for participant orientation, .87 for orderliness, .79 for stimulation, and .82 for the total score. The improvement is substantial, but does it pay off in higher effect sizes? Regrouping all participants trained by the GCT outlier under PCT substantially affects the result—and, most importantly, in the correct theory-derived direction. Attitude toward improving behavior via psychological knowledge and reactions in test situations showed effect sizes of $r = .26$, $t(92) = 2.58$, $p < .02$, and $r = .21$, $t(91) = 2.00$, $p < .05$, favoring PCT over GCT. Communicative sensitivity favored GCT, with an effect size of $r = .30$, $t(90) = 2.95$, $p < .01$. In an area in which nothing seemed to work, we now have effect sizes at least of medium size and in the right direction postulated a priori by program theory. What a difference for summative conclusions!

Testing Broadbent's Theory of Cognitive Control (Klumb, 1995)

The naturalistic approach to cognitive processes has been criticized by some researchers (e.g., Banaji & Crowder, 1989; Rabbitt, 1990) and has been defended by others (e.g., Ceci & Bronfenbrenner, 1991; Reason, 1991). In our view, it is not a question of accepting or rejecting an approach as a whole but of pointing out concrete problems and, when possible, adding some ideas toward their solution. As a case in point, let us look at Broadbent's theory of cognitive control. This theory has been investigated on the basis of different methods, one of which is the Cognitive Failures Questionnaire (CFQ; e.g., Broadbent, Cooper, FitzGerald, & Parkes, 1982). This inventory assesses the subjective frequencies of a wide range of everyday failures of action, memory, and perception that are assumed to have a common basis: an inefficient and inflexible style of attentional resource management.

In an attempt to validate a German version of the CFQ within the domains of everyday performance that are determined by the content universe of its items, Klumb (1995) designed a quasi-experiment. She selected three settings—libraries, dry cleaners, and a lost-property office—in which everyday mental slips and lapses could be observed with particular frequency and their authors could be questioned. The CFQ score of those clients was determined on the basis of the individuals who returned books late, tried to pick up their cleaned clothes without a ticket, or were looking for an object they had lost, respectively. These individuals constituted the experimental groups. Individuals who did not show the behavior in question at the same times and locations were assigned randomly to the control groups. In the lost-property office, these were people who reported to be present on behalf of somebody else. As a manipulation check, individuals within experimental and control groups were asked to indicate how often each of the three target failures (i.e., returning books late, forgetting dry-cleaner's tickets, and losing objects) had happened to them in the last 6 months. Table 10.1 shows the results.

Table 10.1. Distribution of Answers to the Control Questions for Experimental and Control Groups in the Respective Settings

	Hardly ever	Quite rarely	Occasionally	Quite often	Very often
Library groups					
Experimental	1	9	9	12	13
	2.3%	20.5%	20.5%	27.3%	29.5%
Control	19	19	13	6	1
	32.8%	32.8%	22.4%	10.3%	1.7%
Dry-cleaning groups					
Experimental	0	4	4	4	2
		28.6%	28.6%	28.6%	14.3%
Control	15	7	0	0	1
	65.2%	30.4%			4.3%
Lost-property office groups					
Experimental	1	15	0	2	0
	5.6%	83.3%		11.1%	
Control	8	8	1	1	0
		44.4%	44.4%	5.6%	

The manipulation checks in the library and the dry cleaners yielded significant chi-squares, whereas the one in the lost-property office did not. This yielded an overall manipulation that was still significant. Because the manipulation check was significant, the overall correlation between the treatment dummies and the CFQ scores was computed and turned out to be $r_{pb} = .18$, which is highly significant with a sample size of 176! Is that a convincing demonstration of the validity of Broadbent's CFQ? Probably not. Many will echo Walter Mischel's (1968) synthesis that explaining the meager proportion of 3% to 4% of the behavioral variance dispositional variables cannot successfully predict human behavior. What about the reliability of the treatment dummy? Reliability in the library group is .30, in the dry cleaners .46, and in the lost-property office .07. The true correlation between CFQ and behavior is dramatically attenuated. This lack of reliability resulted in a severe loss of power. What about correcting for attenuation or for effects of dichotomizing the continuous variable of failure intensity?

We could use the full information of all continuous ratings and aggregate this information over all three situations, resulting in a treatment intensity variable called MACT_3. Another possibility would be to believe what people said. Those who told us that such a failure happened to them only quite rarely or hardly ever, although they had forgotten their ticket in the specific situation, were reclassified to the control group; that is, they were assigned a score of 0 in the treatment dummy. Those who agreed that such a failure happened to them more often than occasionally, although not having forgotten their ticket in that specific situation, were reclassified to the experimental group (dummy score of 1). This recoded dummy is called CONDNEW. Now we can compute

Table 10.2. Testing Broadbent's Theory of Cognitive Failures With Different Variants of Treatment Operationalization

	CFQSCORE				
	E	COND	CONDNEW	CONDSUM	MACT_3
CFQSCORE	1.000				
COND	0.181	1.000			
CONDNEW	0.372	0.542	1.000		
CONDSUM	0.488	0.318	0.667	1.000	
MACT_3	0.542	0.413	0.612	0.794	1.000

Note. Pearson correlation matrix with original treatment dummy COND, reclassified dummy CONDNEW; CONDSUM is an aggregate over the three condition dummies and MACT_3 is the sum over all original ratings of intensity of cognitive failures in the three situations. Number of observations: 176.

the correlations of these modified treatment variables with the CFQ scores. They are displayed in Table 10.2.

Note that the resulting validity coefficients have climbed from the original .18 to .54 with MACT_3! The variance explained by CFQ is greater than 25% compared with the earlier meager 3% to 4%. What about the credibility and fate of Broadbent's theory? This evaluation is left to you. To be sure, the whole investigation was a quasi-experiment rather than a true experiment. This fact notwithstanding, we were able to demonstrate how we can fool ourselves (and others) in testing theories by not taking into account the reliability of our treatments!

Five-Data-Box Conceptualization and Symmetry: Some Further Promises for Explanation

The synthesis of the Northwestern school of thought with Cronbach's approach, the symmetry principles of the lens model, and a bit of thought about the treatment variables from a psychometricians stance gives some possible explanations for still other problems psychology has faced. Using Cohen's favorite visualization tools—Venn diagrams—allows us to demonstrate how much more power we can bring into designs with that synthesis of both schools (Figure 10.10).

When randomization was successful, the ETR-box variables correlated with neither the NTR-box nor the PR-box variables. This is the major advantage of getting unbiased estimates of the causal effects of the treatment by using the Northwestern path. Yet using variables from all three boxes promises to bring a maximum of power into the design. Selection into treatment is visualized with the overlap of the PR with the NTR box within the CR-box variance. Yet these selection effects can be modeled according to the knowledge about time order.

We have seen in the previous examples that treatment reliability often is very low. This being the case, we can explain another disappointment in psychological and educational research. Cronbach and Snow (1977) looked

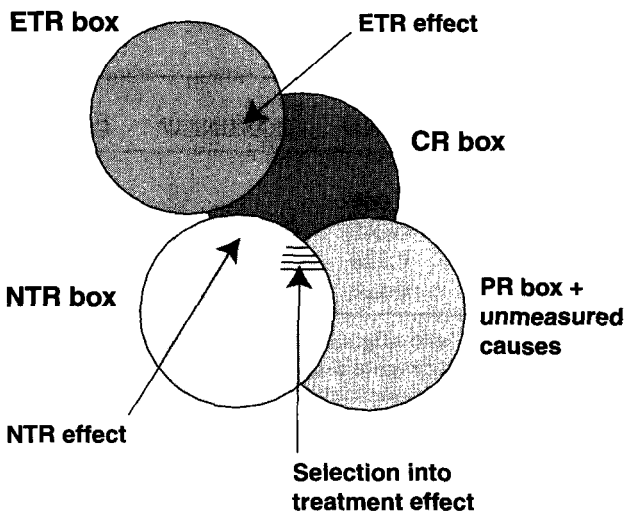


Figure 10.10. Different effects using the five-data-box conceptualization. ETR box = experimental treatment box; CR box = criterion box; PR box = predictor box; NTR box = nonexperimental treatment box.

for aptitude-by-treatment interactions (ATIs), but the overall results of ATI research ended with the depressing summary of Cronbach that interactions were hardly replicable and do not at all generalize. Yet if treatment integrity and therefore its reliability are low, then the reliability of the interaction terms of the partialed product is also low. Aptitude reliability most often is good, but multiplying a variable of poor reliability with one of good reliability still results in an interaction term of mediocre reliability. So should we wonder that interactions did not generalize?[1]

A third promise is a spotlight on the quantitative–qualitative debates. Clinicians often are disappointed that effects they believe they see in their daily practice do not show up after quantification and extensive program evaluation. One can understand that quantification becomes the scapegoat as a consequence. (At the American Evaluation Association now, qualitative interest groups outperform the quantitative ones by a factor of 3 to 4.) The clinicians often check their cases, contrasting them with some matched healthy ones. Although this can be good practice, not being aware of the massive enhancement of range that comes with using such extreme group designs, these individuals easily fell into the trap of overestimating effect sizes. Assume in the context of discovery that they are qualitatively assessing a normally distributed z-score (SD = 1) composite cause and have five cases that are 3 standard deviations above the mean. Suppose they contrast them with five cases 3 standard deviations below the mean. Then their sample standard deviation in z scores is larger than 3, so the quotient u (Figure 10.5) is also greater than 3. The

[1]Werner Wittmann discussed possibilities for reanalysis with the late Dick Snow at Stanford University but owing to his untimely death, it could not be realized.

nomogram tells us what disappointments result once a representative sample is available. What seems to be a medium-sized (.30) effect goes down to a small one, or what was thought to be a large effect (almost .70) changes to a medium-sized one, which, because of the lack of power, might not even be significant.

Finally, a fourth effect is that we might look in the wrong direction when prediction is less than perfect. The case in Figure 10.3b hints at this; we might have already more information than we need for prediction. It is not that something is missing with regard to the criterion. Our predictor contains reliable systematic, but unwanted, variance, which attenuates validity in the same way as random error does. Theory-derived suppressor principles help here and in Figure 10.3d. The appropriate data analysis is set correlation, with its possibilities of partialing unwanted variance (Cohen et al., 2003).

Summary and Conclusions

The synthesis of the Northwestern school of thought concerning basic and applied research with ideas and challenges from its critics paid off, as demonstrated with examples from different areas of research. Similar successes resulted in large-scale evaluation projects in the German health and rehabilitation system (Wittmann, Nübling, & Schmidt, 2002), as well as research about the relationships among working memory, intelligence, knowledge, and complex problem-solving performance in complex computer-based business games (Wittmann & Suess, 1999) not reported here. The key concepts in all reported examples had been the application of symmetry principles in relating predictors, causes, and effects. Of special additional importance was incorporating psychometric principles into the experimental treatment to improve its measurement and to shed light into the black box. Investing more in the assessment of criteria and taking a larger slice out of human behavior over longer periods helped as well. We are reminded that time-series designs are the strongest quasi-experimental ones in terms of internal validity. Tools coined as ambulatory assessment have been developed to better assess behavior, feelings, emotions, and performance in real-life field settings. Fahrenberg and Myrtek (1996, 2001) contributed to their development and described the potential and promises. We are confident that assessment, measurement, theory testing, and construct validation will reach new horizons by integrating these tools into our research designs.

Epilogue and a Personal Note

It is a great pleasure to have Lee Sechrest, the "Method Man," with his rigorous Northwestern roots and background, as a role model. His ideas about measurement and hints at neglected problems of treatment strength and integrity stimulated our own thinking. We have been impressed by the breadth and the sheer number of areas in which he did research and consultation. We tried to follow his footsteps in psychotherapy, clinical psychology, personality, health, program evaluation, and evaluation research but could hardly keep pace. We

are grateful for more than a decade of exchanging ideas, as well as students and coworkers. We enjoyed his regular visits to Germany and the many symposia at international conferences he helped organize. We are grateful for the time he shared with us and especially for his invitations to the famous EGAD (Evaluation Group for the Analysis of Data) dinners at these meetings.

References

American Journal of Evaluation. (2003). Historical records. *American Journal of Evaluation, 24*, 261–288.

Banaji, M. R., & Crowder, R. G. (1989). The bankruptcy of everyday memory. *American Psychologist, 44*, 1185–1193.

Boruch, R. F., & Gomez, H. (1977). Sensitivity, bias, and theory in impact evaluations. *Professional Psychology, 8*, 411–434.

Broadbent, D. E., Cooper, P. F., FitzGerald, P., & Parkes, K. R. (1982). The Cognitive Failures Questionnaire (CFQ) and its correlates. *British Journal of Clinical Psychology, 21*, 1–16.

Brunswik, E. (1955). Representative design and probabilistic theory in functional psychology. *Psychological Review, 62*, 236–242.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Cattell, R. B. (1988). The data box: Its ordering of total resources in terms of possible relational systems. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 69–130). New York: Plenum Press.

Ceci, S. J., & Bronfenbrenner, U. (1991). On the demise of everyday memory: "The rumors of my death are much exaggerated" (Mark Twain). *American Psychologist, 46*, 27–31.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65*, 145–153.

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin, 70*, 426–443.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Rev. ed.). New York: Academic Press.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7*, 249–253.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.

Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). London: Erlbaum.

Cook, T. D. (1993). A quasi-sampling theory of the generalization of causal relationships. In L. Sechrest & A. G. Scott (Eds.), *New directions for program evaluation: Understanding causes and generalizing about them* (Vol. 57, pp. 39–82). San Francisco: Jossey-Bass.

Cook, T. D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis, 24*, 175–199.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Boston: Houghton-Mifflin.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12*, 671–684.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30*, 116–127.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs.* San Francisco: Jossey-Bass.

Cronbach, L. J., Ambron, S., Dornbusch, S., Hess, R., Hornik, R., Phillips, D., et al. (1980). *Toward reform of program evaluation*. San Francisco: Jossey-Bass.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York: Irvington.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*, 95–106.

Epstein, S. (1980). The stability of behavior: II. Implications for psychological research. *American Psychologist, 35*, 790–806.

Epstein, S. (1983). Aggregation and beyond: Some basic issues on the prediction of behavior. *Journal of Personality, 51*, 360–392.

Epstein, S., & O'Brien, E. J. (1985). The person–situation debate in historical and current perspective. *Psychological Bulletin, 98*, 513–537.

Fahrenberg, J., Hampel, R., & Selg, H. (2001). *Freiburger Persönlichkeitsinventar FPI-R* [Freiburger Personality Inventory] (7th ed.). Göttingen, Germany: Hogrefe.

Fahrenberg, J., & Myrtek, M. (Eds.). (1996). *Ambulatory assessment: Computer-assisted psychological and psychophysiological methods in monitoring and field studies*. Göttingen, Germany: Hogrefe & Huber.

Fahrenberg, J., & Myrtek, M. (Eds.). (2001). *Progress in ambulatory assessment: Computer-assisted psychological and psychophysiological methods in monitoring and field studies*. Seattle, WA: Hogrefe & Huber.

Fahrenberg, J., Myrtek, M., Kulick, B., & Frommelt, P. (1977). Eine psychophysiologische zeitreihenstudie an 20 studenten über 8 wochen [A psychophysiological longitudinal study of 20 students over 8 weeks]. *Archiv für Psychologie, 129*, 242–264.

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research*. Reading, MA: Addison-Wesley.

Glass, G. V. (1983). Evaluation methods synthesized: Review of L. J. Cronbach designing evaluations of educational and social programs. *Contemporary Psychology, 28*, 501–503.

Hammond, K. R. (Ed.). (1966). *The psychology of Egon Brunswik*. New York: Holt, Rinehart & Winston.

Hammond, K. R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice*. New York: Oxford University Press.

Hammond, K. R., & Stewart, T. R. (Eds.). (2001). *The essential Brunswik: Beginnings, explications, applications*. New York: Oxford University Press.

Hilgard, E. R. (1955). Discussion of probabilistic functionalism. *Psychological Review, 62*, 226–228.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

Kerlinger, F. N. (1973). *Foundations of behavioral research*. London: Holt, Rinehart & Winston.

Klumb, P. L. (1995). Cognitive failures and performance differences: Validation studies of a German version of the Cognitive Failures Questionnaire. *Ergonomics, 38*, 1456–1467.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage.

Lipsey, M. W. (1993). Theory as method: Small theories of treatments. In L. Sechrest & A. G. Scott (Eds.), *New directions for program evaluation: Understanding causes and generalizing about them* (Vol. 57, pp. 5–38). San Francisco: Jossey-Bass.

Lösel, F., Köferl, P., & Weber, F. (1987). *Meta-Evaluation der Sozialtherapie* [Meta-evaluation of social therapy]. Stuttgart, Germany: Enke.

Lösel, F., & Wittmann, W.W. (1989). The relationship of treatment integrity and intensity to outcome criteria. In R. F. Conner & M. Hendricks (Eds.), *New directions for program evaluation: Vol. 42. International innovations in evaluation methodology* (pp. 97–107). San Francisco: Jossey-Bass.

Matt, G. E. (2003). Will it work in Münster? Meta-analysis and the empirical generalization of causal relationships. In R. Schulze, H. Holling, & D. Böhning (Eds.), *Meta-analysis: New developments and applications in medical and social sciences* (pp. 113–139). Göttingen, Germany: Hogrefe & Huber.

Mischel, W. (1968). *Personality and assessment*. New York: Wiley.

Mischel, W., & Peake, P. K. (1982). Beyond déjà vu in the search for cross-situational consistency. *Psychological Review, 89*, 730–755.

Rabbitt, P. (1990). Age, IQ, and awareness, and recall of errors. *Ergonomics, 33*, 1291–1305.

Reason, J. (1991). Self-report questionnaires in cognitive psychology: Have they delivered the goods? In A. Baddeley & L. Weiskrantz (Eds.), *Attention: Selection, awareness and control* (pp. 406–423). Oxford, England: Oxford University Press.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. Cambridge, England: Cambridge University Press.

Rossi, P. H. (1978). Issues in the evaluation of human services delivery. *Evaluation Quarterly, 2*, 573–599.

Ryans, D. G. (1960). *Characteristics of teachers*. Washington, DC: American Council on Education.

Sechrest, L. (1986). Modes and methods of personality research. *Journal of Personality, 54*, 318–331.

Sechrest, L., West, S. G., Phillips, M. A., Redner, R., & Yeaton, W. H. (1979). Some neglected problems in evaluation research: Strength and integrity of treatments. In L. Sechrest (Ed.), *Evaluation studies review annual* (Vol. 4, pp. 15–38). Beverly Hills, CA: Sage.

Sechrest, L., Schwartz, R. D., Webb, E. J., & Campbell, D. T. (1999). *Unobtrusive measures*. Newbury Park, CA: Sage.

Sechrest, L., & Yeaton, W. H. (1981). Assessing the effectiveness of social programs: Methodological and conceptual issues. *New Directions for Program Evaluation, 9*, 41–56.

Sechrest, L., & Yeaton, W. H. (1982). Magnitudes of experimental effects in social sciences research. *Evaluation Review, 6*, 579–600.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*, 309–316.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized inference*. Boston: Houghton Mifflin.

Smith, M., Glass, G. V., & Miller, T. I. (1980). *The benefits of psychotherapy*. Baltimore: Johns Hopkins University Press.

Spinner, M. (1991). *Replikation der Meta-Analyse deutschsprachiger Psychotherapieeffektforschung der Jahre 1971–1982 unter Integration unberücksichtigter und neuerer Arbeiten der Jahre 1982–1988* [Replication of the meta-analysis of German-speaking psychotherapy effect size research 1971–1982 considering new research from 1982 to 1988]. Unpublished master's thesis, University of Freiburg, Germany.

Suchman, E. A. (1967). *Evaluative research: Principle and practice in public service and social action programs*. New York: Russell Sage Foundation.

Thorndike, R. L. (1949). *Personnel selection: Test and measurement techniques*. New York: Wiley.

Tucker, L. R. (1964). A suggested alternative formulation in the developments by Hursch, Hammond & Hursch; and by Hammond, Hursch & Todd. *Psychological Review, 71*, 528–530.

Wittmann, W. W. (1985). *Evaluationsforschung: Aufgaben, Probleme und Anwendungen* [Evaluation research: Tasks, problems and applications]. Berlin, Germany: Springer-Verlag.

Wittmann, W. W. (1987a). Grundlagen erfolgreicher forschung in der psychologie [Foundations of successful research in psychology]. *Diagnostica, 33*, 209–226.

Wittmann, W. W. (1987b). Meta-analysis of German psychotherapy outcome studies: The importance of research quality. In W. Huber (Ed.), *Progress in psychotherapy research* (pp. 770–787). Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain.

Wittmann, W. W. (1988). Multivariate reliability theory: Principles of symmetry and successful validation strategies. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 505–560). New York: Plenum Press.

Wittmann, W. W. (2002). Brunswik-Symmetrie: Ein Schlüsselkonzept für erfolgreiche psychologische Forschung [Brunswik symmetry: A key concept for successful psychological research]. In M. Myrtek (Ed.), *Die Person im biologischen und sozialen Kontext* [The person in a biological and social context] (pp. 163–186). Göttingen, Germany: Hogrefe.

Wittmann, W. W., & Matt, G. E. (1986). Meta-Analyse als Integration von Forschungsarbeiten am Beispiel deutschsprachiger Arbeiten zur Effektivität von Psychotherapie [Meta-analysis as an integration of research exemplified for German studies on the effect of psychotherapy]. *Psychologische Rundschau, 37*, 20–40.

Wittmann, W. W., & Suess, H.-M. (1999). Investigating the paths between working memory, intelligence, knowledge, and complex problem-solving performances via Brunswik-Symmetry.

In P. L. Ackerman, P. C. Kyllonen, & R. D. Roberts (Eds.), *Learning and individual differences. Process, trait, and content determinants* (pp. 77–108). Washington, DC: American Psychological Association.

Wittmann, W. W., & Walach, H. (2002). Evaluating complementary medicine: Lessons to be learned from evaluation research. In G. Lewith, W. B. Jonas, & H. Walach (Eds.), *Clinical research in complementary theories, problems and solutions* (pp. 98–108). London: Churchill Livingston.

Wittmann, W. W., Nübling, R., & Schmidt, J. (2002). Evaluationsforschung und programmevaluation im gesundheitswesen [Evaluation research and program evaluation in health care]. *Zeitschrift für Evaluation, 1*, 39–60.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research, 10*, 557–585.

Yeaton, W. H., & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology, 49*, 156–167.

Zee, A. (1989). *Fearful symmetry. The search for beauty in modern physics*. New York: MacMillan.