Running head: HOW TO FOOL YOURSELF

Wittmann, W. W., & Klumb, P. L. (2006). How to fool yourself with experiments in testing psychological theories. In R. R. Bootzin &P. E. McKnight (Eds.), Strengthening Research Methodology: Psychological Measurement and Evaluation. Washington, DC: American Psychological Association.

See: http://www.apa.org/books/4318026t.html

How to fool yourself with experiments in testing theories in psychological research

Werner W. Wittmann

Universität Mannheim, Germany

Petra L. Klumb

Technische Universität Berlin, Germany

1. Legacy and impact of the Northwestern school

The Northwestern school as Glass (1983) has coined it, no longer resides in the Department of Psychology at Northwestern University, Evanston, IL, but its members are spread over the whole country 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 which have to be studied, learned and digested by every student worldwide, who seriously wants to do research in social sciences. The Northwestern schools 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 quasiexperiments, especially in the context of education the field most resistant to experimentation. Cook (2002) analyzes these reasons of 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 quasiexperimental designs, but their 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 placed 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, which allows us given certain circumstances and background restriction to choose between different approaches.

2. The five data box conceptualization a comprehensive framework for research and program evaluation

For this purpose we have developed a framework coined the five data box conceptualization (Wittmann, 1985, 2002, Wittmann and Walach, 2002), which refers to five different sources of information one must consider and gather in the process of basic or applied research. Fig. 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 its three dimension: subjects, variables and situations/time (Cattell, 1988).

Fig.1 about here

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. Stakeholders are subjects interested in the results; i.e. 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 considers ones peers, because difficulties result when there is a lack of interest in the topic which 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 and to control the status before research and to answer any questions about selection effects as regards to the population of interest. The ETR-box maps the actively manipulated treatment variables and as subjects the members of the randomized experimental and control groups. In ANOVA parlance, these are the independent variables called fixed or random factors and their interactions. The NTR-box contains all treatment aspects, which could not be randomized, e.g. 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 onto the ETR-box and the PR-box we follow the Northwestern path. Regressing the CR-box onto the NTR-box and the PR-box, we follow the Southwestern path. The wind rose at the upper left corner of Fig.1 serves as a guide to read 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) coined 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, assessment methods and decision aids.

Lee Sechrest has contributed to assessment (Sechrest, 1986, Sechrest, Schwartz, Webb, and Campbell, 1999), to debates about quantitative vs. qualitative research and the problems related to treatment integrity, fidelity, implementation and strength (Sechrest, Phillips, Redner, and Yeaton, 1979, Sechrest and Yeaton, 1981, 1982, Yeaton and Sechrest, 1981). Lack of treatment integrity or failures in implementing a program can easily explain why a 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 fifties, we find a heated debate between proponents of the experimental and those of the representative design. Egon Brunswik (1955) who proposed the representative design was heavily criticized especially from 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 spuriousness. 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, quasiexperimental or plain correlational can be analyzed by the general linear model, i.e. multiple regression/correlation. His paper was expanded into a full textbook (Cohen and Cohen, 1975), which has seen its third edition (Cohen, Cohen, West & Aiken, 2003).

The four right-most boxes are related with directed arrows mapping the time paths between them. Only the relationship between the ETR and the NTR- box is denoted with a double-headed arrow indicating the gradual decline from a fully randomized design to more quasi-experimental and correlational one. Interestingly the title of Cook and Campbell (1979) already was ”Quasi-Experimentation...” demonstrating that the Northwestern school was fully aware of the problems 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.

3. Brunswik-symmetry a key concept for successful psychological research

Looking for reasons why 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) describes 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 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 and Stewart, 2001). We have focused on his lens model and have used it to look at the relationship between our data boxes. Fig.2 visualizes e.g. the PR-CR-box relationship.

Fig. 2 about here

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 we can 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. Fig.3 distinguishes four variants of asymmetry.

Fig. 3a, b, c and d about here

Fig. 3 a shows the case of full asymmetry, which is the case where 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 predictive validity. This case happens by choosing assessment according to their psychometric reliability only and not in terms of their construct relevance or as we coin them their construct reliability. Nevertheless, it is an interesting case because according to Campbell and Fiske's (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. Fig.3 b denotes the case where 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 and O'Brien, 1985) focused on the importance of aggregation and demonstrated that he could boost on validity, but Mischel (Mischel and Peake, 1982) insisted on the predictability of behavior in the specific situation.

Fig. 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 loosing 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. Fig. 3 d is the hybrid case, where we have a mismatch at the same level of generality, i.e. only partial overlap. Validity is different from zero but is this indication of convergent or of discriminant validity?

This visualization is immediately evident and it is easily to find examples where 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. Fig. 4 shows the ETR-box.

Fig. 4 about here

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 to 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 the 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 we invested so much in measuring the dependent, but forgot to do so for the independent in experiments. What insights result if we look at the independent 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 plus variance within groups. Looking at the treatment/control dummy (Fig.4) we immediately see that the pooled variance within groups is zero. Thus, an experimenter implicitly assumes that the reliability of the independent is always one! But this is wishful thinking, due to compliance, implementation, John Henry 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 to be discussed below, 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 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 also via a mathematical numerical equation, thanks to an elegant solution given by Tucker (1964). Eq. 1 shows the original form of that equation:

$$r_{PR,CR} = G_{PR,CR} R_{PR} \cdot R_{CR} + C_{PR,CR} \sqrt{\left(1 - R_{PR}^2\right)\left(1 - R_{CR}^2\right)} \tag{1}$$

The observed predictor/criterion correlation is explained by several parameters. The first part is related to a linear model and the second one to a model, which 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 onto 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 partialled. In developing that equation 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 due to selection effects and sampling error, so we simply augmented these concepts into Tucker's lens-model equation. Eq. 2 shows this augmented equation for the relationship between the ETR- and the CR- box because our focus here is on how we can fool ourselves with experiments.

$$\begin{aligned} r_{ETR,CR}^{observed} &= S_l \sqrt{r_{tt}^{ETR(l)} \cdot r_{tt}^{CR(l)}}\; G_{ETR(l),CR(l)}^{true} \cdot R_{ETR(l)} \cdot R_{CR(l)} + \\ & S_n \sqrt{r_{tt}^{ETR(n)} \cdot r_{tt}^{CR(n)}}\; C_{ETR(n),CR(n)}^{true} \cdot R_{ETR(n)} \cdot R_{CR(n)} + e \end{aligned} \tag{2}$$

The additional parameter are as follows: $r_{tt}^{ETR(l)}$ and $r_{tt}^{CR(l)}$ the classical psychometric reliabilities of the linear models of the operationalization of the experimental treatment and the criterion respectively. The terms $r_{tt}^{ETR(n)}$ and $r_{tt}^{CR(n)}$ are the psychometric reliabilities of the nonlinear models and e stands for error. S_l and S_n meaning linear and nonlinear ones denoting selection effects. Dawes and Corrigan (1974) have 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 only equal to one, when the sample sd is equal to the population sd, when sd_{sample} is smaller than sd_{pop}, S is smaller than one and when sd_{sample} is larger than sd_{pop} S turns out to be larger than one. S is only a placeholder to denote the selection problems that are known since Thorndike (1949). Hunter and Schmidt (1990) give the following equation:

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

Here in (3) is a typo, the symbol for division after r_{pop} is missing, namely r_{sample} = u r_{pop} / V(u²-1)r_{pop} +1

(4) $$u = sd_{sample} / sd_{pop}$$

To demonstrate how large S gets under selection, we have constructed a nomogram for a rough calculation of these effects (Fig.5).

Fig.5 about here

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 sd in the sample is only half of the sd in the population, i.e. u=. 50 with a large effect size we get only a sample effect size of r =. 28 and S would be .28/. 50= .56. If u= 2.0 than sample r 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 Fig. 6

Fig.6 about here

The true effect size in the population is surrounded by parameters that either attenuate or augment it. There are six dangers to underestimate a true effect and only two dangers to overestimate it. Therefore, the odds of underestimation are higher than 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, Hunter and Schmidt, and Rosenthal among many others have contributed to popularizing meta-analysis. Glass synthesized experiments in psychotherapy, Hunter and Schmidt started in synthesizing validity coefficients in personnel selection research and coined their approach as validity generalization, and Rosenthal synthesized the p-values from significance testing. All these approaches are now under a common framework see Rosenthal, Rosnow, and Rubin (2000). The d- and r- families of effect sizes easily can be transformed into one another. The effect size r can be transformed into Cohen’s d as follows:

(5) $$d = r / \mathrm{sqrt}\left(pq\left(1 - r^2\right)\right)$$

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

For the experimental approach, we must reflect what the possible distribution of the independent is. Is it normally distributed, rectangular or something else? Causes do not have a distribution they only differ in dosage level or strength. 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) discusses that independent variable and the role of theory. He distinguishes 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, second and third nonlinear functions mapping effects for strong or weak doses, fourth and fifth U-shaped and inverted-U functions. These theoretical considerations are very important in realizing the MAXMINCON principles recommended by Kerlinger (1973), which state that one should maximize the effect between groups but minimize the variance within and control for unwanted systematic variance. The experimental and control group must differ in the dosage level, and the split in which we map our treatment dummy must correspond to that level where we assume an onset of the response occurs. For such unitary causes we need a lot of theoretical knowledge on where to make the split. In most program evaluation whole treatment packages are the interventions, we can assume that several causes should be at work. Whatever the dose response functions of the unitary causes are the composite causes 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 to the MAX- principle brings us back to the problems of enhancement of range mentioned above. 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 knowing parameter S we can correct the effect we find in such designs once 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) has 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 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 Sedlmaier and Gigerenzer (1989) reported their second look at research results.

4. Meta-analysis and the Brunswik lens-model equation

Hunter and Schmidt (1990) used the parameters of Fig. 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 seventy-five percent of the variance of observed effect sizes can be explained by these artifacts then the overall effect can be generalized and there is no need for looking additionally at moderators which can explain the true effect variability. They used this 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 in each selection situation anew! 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. He 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 and 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) where analyzed, there were also 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, 1987a). Behavioral interventions had higher effect sizes compared to 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 check lists 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 with Fig. 3c. A lead indicator was whether the design a-priori was designed as a follow-up study, taking a larger slice of the time/situation coordinate of the CR-box. Those who did had better hypothesis about the stability of effects, their generalizability over time, used multi-method and multivariate assessments, focused more on specific aspects of personality and specific subgroups. One can speculate when a follow-up design with extended post measures over time is used the researchers already have accumulated more knowledge about causal effects making them confident that the intervention works, otherwise they wouldn't have invested the extra resources these designs require.

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

To test the Brunswik symmetry principles we built an index mapping symmetry between the causes and effects in terms of external and construct validity, low scores indicating high symmetry and high scores higher asymmetry. Fig. 7 shows effect size box plots as a function of asymmetry and the overall distribution bolsters our hypothesis.

Fig. 7 about here

5. 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 eight weeks, which we used as an attempt to validate Eysenck's personality theory (Wittmann, 1987b). The second is a program evaluation study of Lösel, Köferl and Weber (1987) about the training effects of prison officers (Lösel and Wittmann, 1989) and the third 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.

5.1 The promise of longitudinal designs for personality traits (Fahrenberg et al., 1977)

Fahrenberg's lab at University of Freiburg is most well known for its focus on psychophysiology. Fahrenberg also developed the most used German speaking personality inventory the "Freiburger Persönlichkeitsinventar (FPI)". The FPI (Fahrenberg, Hampel, and Selg, 2001) among other dimensions also measures Eysenck's extraversion and emotional lability (neuroticism) factors. In the study we assessed twenty students over eight weeks. At the beginning, they took the FPI and over the two-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 got ratings 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 the Fahrenberg et al. study. From a theoretical stance, we assumed that traits are dispositional constructs. A disposition is a tendency only 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 between single act, repeated single act, and multiple acts in a relatively specific situation or timeframe and repeated multiple act criteria, which aggregate functionally equivalent behaviors and feelings (RMAC) over many situations or periods. For the extraversion RMAC, we could aggregate over 60 days. The RMAC for emotional lability was constructed via 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 eight 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. Fig.8 shows the results:

Fig. 8 about here

Applying Campbell and Fiske's (1959) principles of convergent and discriminant validity the results are impressive. Eysenck's theory postulates E and N 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 using the reliability estimates for correction for attenuation. Although we are aware of the limitations of a sample size of N=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 to a possible solution of the Epstein/Mischel debate. Personality traits might be very good predictors for aggregated multiple act criteria but not so well 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, i.e. 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 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 in order to change them we need a corresponding symmetrical intervention, which can only be a multifaceted treatment package. We saw that the variability in alcohol, drug and medication plays a role. It was not the mean level in these facets but their ups and downs, so what triggers their onset? How should we deal with relapse prevention? How can we stabilize the 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 the subjects regard them as a problem.

5.2 Training prison officers with psychological interventions (Lösel, Köferl and Weber, 1987)

Prison officers are the persons 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 behavioral therapy background and four trainers with a Rogerian background were used. They were partially randomized and matched to train and educate eleven or twelve prison officers in each group. These groups were compared to each other and to an untrained control group. The program-centered groups (PCT) followed the tradition of behavioral learning theory, whereas the group-centered training (GCT) followed the tradition of T-group laboratories. As criterion measures theory derived outcome variables were chosen to map effects, which can be best expected based on what each intervention trains. Attitudes towards 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 trained in the PCT groups and the latter closer to what was trained in the GCT groups. The training took one full week, all training sessions were videotaped and the post-tests were given five month after training. Data analysis showed no significant differences between PCT and GCT on the first criterion. The effect size was r=. 11 (t= 1.08, df= 92). In the second one the effect size was r= .06 not significant (t= 0.53, df = 91) and neither to the control group. For the third one most relevant for GCT there was a significant difference to the control group but no significant one to PCT. Effect size here was again r= .11 (t= 1.09, df =91). The summative evaluation would have ended as another example of no difference research or an additional study to bolster Rossi's 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 for 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 Fig. 9

Fig. 9 about here

As can be seen PCT is rated more homogenous and with higher average intensity on all three dimensions. Within GCT, 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 MAX- principle was not successfully established, treatment homogeneity within groups was lacking. As Fig. 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 subjects trained by the GCT outlier under PCT substantially affects the result and most importantly in the correct theory derived direction. Attitude towards improving behavior via psychological knowledge and reactions in test situations showed effect sizes r= .26 (t= 2.58, df= 92, p< .02) and r= .21 (t= 2.00, df = 91, p< .05) favoring PCT over GCT. Communicative sensitivity favored GCT with an effect size of r= .30 (t = 2.95, df = 90, p < .01). In an area where 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!

5.3 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 towards 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 based on 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 loosing objects) had happened to them in the last six months. Table 1 shows the results.

Table 1 about here

The manipulation checks in the library and the dry cleaner yielded significant Chi-squares while the one in the lost property office did not. This yielded an overall manipulation that was still significant. Since 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 3-4% 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 cleaner .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 only happened to them quite rarely or hardly ever, although they had forgotten their ticket in the specific situation, are reclassified to the control group, i.e., were assigned a score of zero 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, are reclassified to the experimental group (dummy score of one). This re-coded dummy is called CONDNEW. Now, we can compute the correlations of these modified treatment variables with the CFQ scores. They are displayed in Table 2.

Table 2 about here

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 to the meager 3-4%. 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!

6. 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 thinking about the treatment variables from a psychometricians stance gives some possible explanations for still other problems psychology has dealt with. Using Cohen's favorite visualization tools ballantines allows us to demonstrate how much more power we can bring into designs with that synthesis of both schools (Fig.10).

Fig.10 about here

When randomization was successful the ETR-box variables do neither correlate with NTR- nor PR-box variables. This is the major advantage to get unbiased estimates of the causal effects of the treatment using the Northwestern path. But 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. But these selection effects can be modeled according to the knowledge about time order.

We have seen in the examples above 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 for aptitude x treatment interactions, but the overall results of ATI-research ended with the depressing summary of Cronbach; that interactions were hardly replicable and they do not generalize. But if treatment integrity and therefore its reliability is low, consequently the reliability of the interaction terms of the partialled product is also low. Aptitude reliability most often is good, but multiplying a variable of low reliability with one of good reliability still results in an interactions 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 to 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 AEA now qualitative interest groups outperform the quantitative ones by a factor of 3 to 4). They often check their cases contrasting them with some matched healthy ones. Although this can be good practice not being aware of massive enhancement of range 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-scored (sd=1) composite cause and have 5 cases which are 3 sds above the mean and they contrast them with 5 cases 3 sds below the mean, then their sample sd in z-scores is larger than 3, so the quotient u (Fig. 5) is also greater then 3. The 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 due to the lack of power might not even be significant.

Finally a fourth derivation is that we might look in the wrong direction when prediction is less than perfect. The case in Fig 3b hints to this, we might have already more information than we need for prediction. It is not that something is missing as regards the criterion. Our predictor contains reliable systematic but unwanted variance which attenuates validity in the same way as random error. Theory derived suppressor principles help here and in Fig. 3d. The appropriate data-analysis is set correlation with its possibilities of partialling unwanted variance (Cohen et al. 2003).

7. 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 and Schmidt, 2002) as well as research about the relationship between working memory, intelligence, knowledge and complex problem solving performance in complex computer-based business games (Wittmann and 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, taking a larger slice out of human behavior over longer time 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) have contributed to their development and describe 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.

8. 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 to neglected problems of treatment strength and integrity stimulated our own thinking. We have been impressed by the breadth and the sheer number of the 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 common symposia at International conferences he helped organizing. We are grateful for the time he shared with us and especially for his invitations to the famous EGAD-dinners at these meetings. Thank you Lee!

References

American Journal of Evaluation (2003). Historical Records. *American Journal of Evaluation, 24(2)*, 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, IL: 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 mulitple regression/correlation analysis for the behavioral sciences*. Hillsdale: Lawrence Erlbaum.

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

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 Publishers.

Cook, T. D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offerd 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, MA: 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., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York: Irvington.

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.

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., & Myrtek, M. (Eds.). (1996). *Ambulatory assessment. Computer-assisted psychological and psychohysiological methods in monitoring and fields studies*. Göttingen: Hogrefe & Huber Publishers.

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

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

Fahrenberg, J., Myrtek, M., Julick, 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: Sage Publ.

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. B. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park: Sage.

Lipsey, M. B. (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 Publishers.

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

Lösel, F. & Wittmann, W.W. (1989). The relationship of treatment integrity and intensity to outcome criteria. In R.F. Conner & M. Hendricks (Eds.). *International innovations in evaluation methodology. New Directions for Program Evaluation, 42* (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: 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.* Oxford: Oxford University Press.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research. A correlational approach*. Cambridge, UK: 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 D.C.: American Council on Education.

Sechrest, L. (1986). Modes and methods of personality research. *Journal of Personality, 54(1)*, 318-331.

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.

Sechrest, L., 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). Beverly Hills, CA: Sage.

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

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 Co.

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

Spinner, M. (1991). *Replikation der Meta-Analyse deutschsprachiger Psychotherapieeffekforschung 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 1982-1988)* Unveröffentlichte Diplomarbeit: Universität Freiburg (unpublished Master thesis University of Freiburg, Germany).

Suchman, E. A. (1967). *Evaluative research: Principle and practice in public service and social action programs*. New York: Russel 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: Springer-Verlag.

Wittmann, W. W. (1987a). 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: Presses Universitaires de Louvain.

Wittmann, W. W. (1987b). Grundlagen erfolgreicher Forschung in der Psychologie [Foundations of successful research in psychology: Multimodal assessment, multiplism, multivariate reliability and validity theory]. *Diagnostica*, *33,* 209-226.

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* (pp. 163-186). Göttingen: 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 D.C.: APA-Books.

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: Chruchill 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: MacMillian.

Author Note

Thanks to Guido Makransky and Tobias Bothe for helping in grammar and style

Prof. Dr. Werner W. Wittmann

Universität Mannheim

Lehrstuhl Psychologie II

Schloß, EO

D-68131 Mannheim

Fon: +49-621-181-2138

Fax: +49-621-181-2129

email: wittmann@tnt.psychologie.uni-mannheim.de

Dr. Petra L. Klumb

Research Group GERM

TU Berlin, BH8

Ernst-Reuter-Platz 1

D-10587 Berlin

Fon: +49-30-314-79429

http://www.tu-berlin.de/fb8/nwg

email: petra.klumb@tu-berlin.de

Footnotes

[1] The senior author discussed possibilities for reanalysis with late Dick Snow at Stanford but due to his untimely death it could not be realized.

Figure Caption

Table 1. Distribution of answers to the control questions for experimental and control groups in the respective settings

Table 2. Testing Broadbent’s theory of cognitive failures with different variants of treatment operationalization

Figure 1. The five data-box conceptualization

Figure 2. The true Brunswik-symmetrical latent structure of nature

Figure 3a. Full asymmetry - The case of nothing works

Figure 3b. Asymmetry due to a broad higher-level predictor

Figure 3c. Asymmetry due to a narrower lower level predictor

Figure 3d. The hybrid case of asymmetry

Figure 4. A closer look at the experimental treatment box

Figure 5. Nomogram for selection effects: Parameter S

Figure 6. The Brunswik-lens-equation for relating experimental treatment (ETR) to criteria (CR)

Figure 7. German psychotherapy effects as a function of symmetry

Figure 8. Testing Eysenck’s E-/N-theory in the Brunswik-symmetry framework

Figure 9. Behavior of group trainers as perceived in single courses (plain lines) and on the average (dotted lines)

Figure 10. Different effects using the five-data-box conceptualization

Table 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 2,3 %	9 20,5 %	9 20,5 %	12 27,3 %	13 29,5 %
Control	19 32,8 %	19 32,8 %	13 22,4 %	6 10,3 %	1 1,7 %
Dry cleaning groups:					
Experimental	0	4 28,6 %	4 28,6 %	4 28,6 %	2 14,3 %
Control	15 65,2 %	7 30,4 %	0	0	1 4,3 %
Lost property office groups:					
Experimental	1 5,6 %	15 83,3 %	0	2 11,1 %	0
Control	8	8 44,4 %	1 44,4 %	1 5,6 %	0

Table 2.

Testing Broadbent’s Theory of Cognitive Failures with Different Variants of Treatment Operationalization

	CFQSCORE	**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

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

Numbers of Observations: 176

Figure 1.

The Five Data-Box Conceptualization

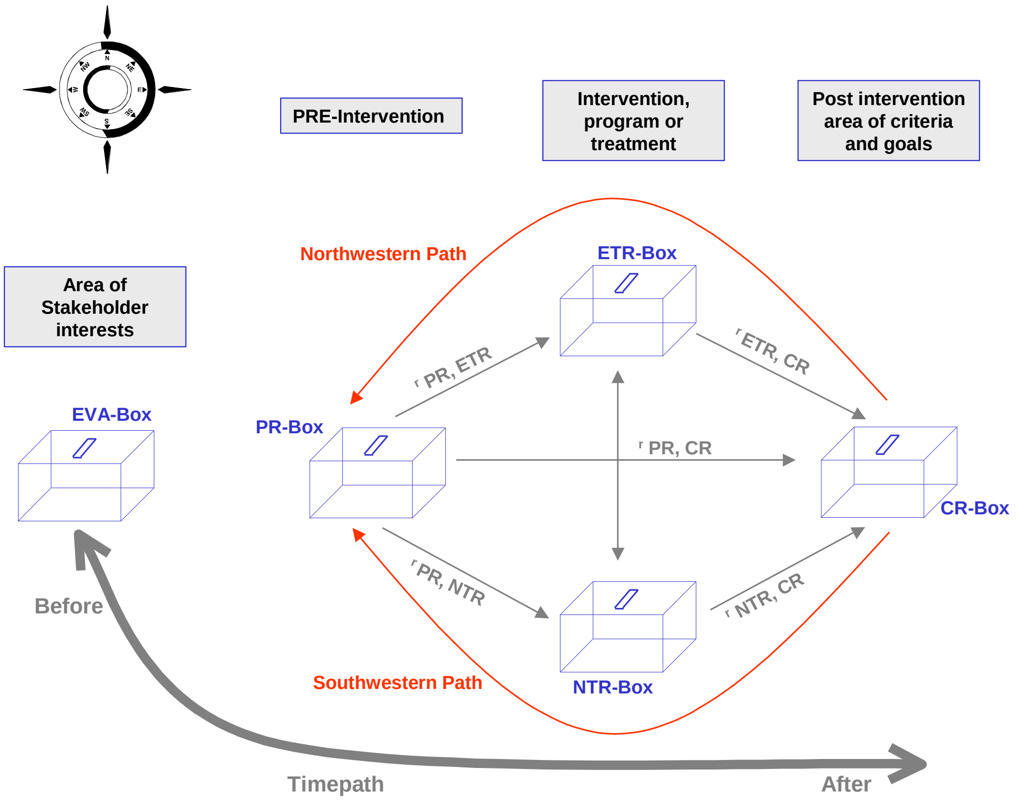


Figure 2.

The true Brunswik-symmetrical latent structure of nature

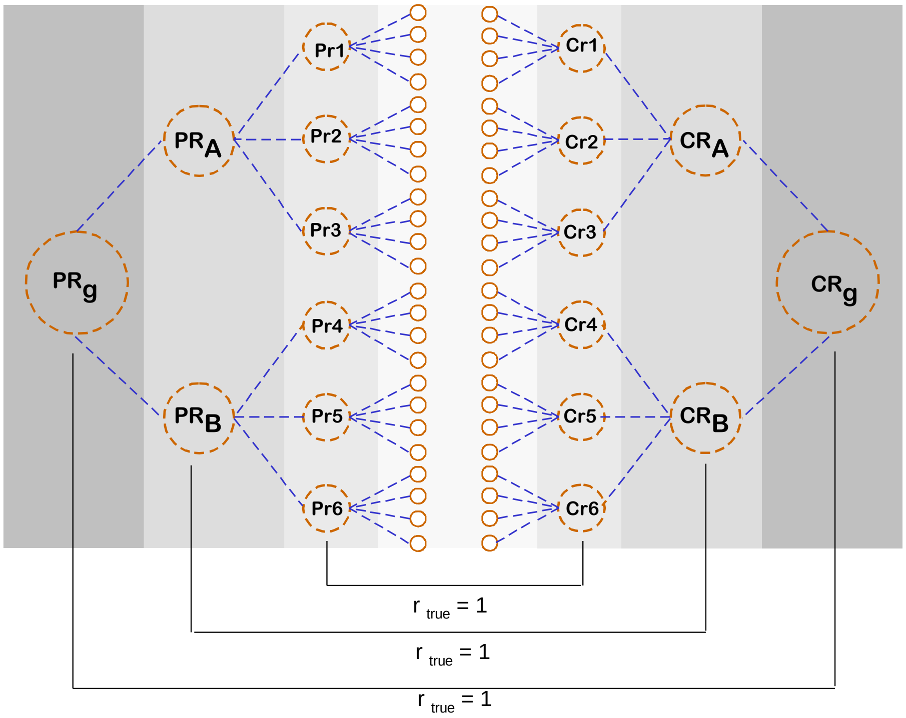


Figure 3a.

Full asymmetry - The case of nothing works

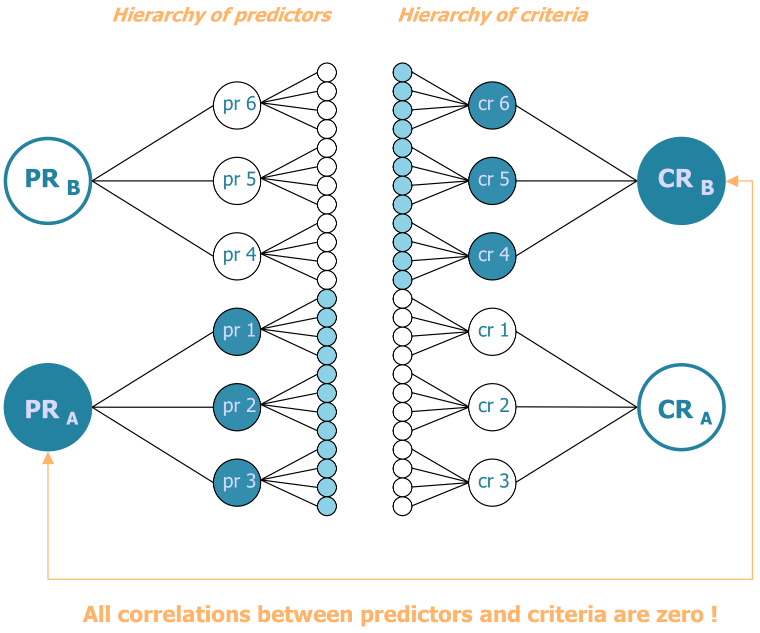


Figure 3b.

Asymmetry due to a broad higher level predictor

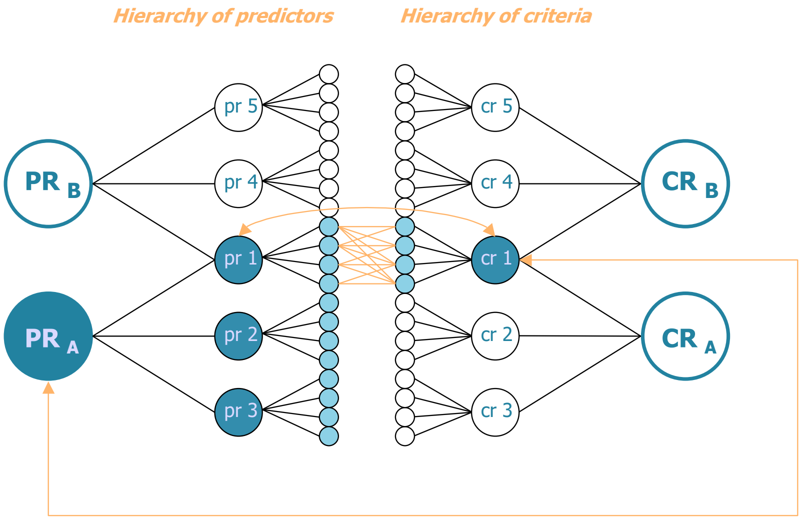


Figure 3c.

Asymmetry due to a narrower lower level predictor

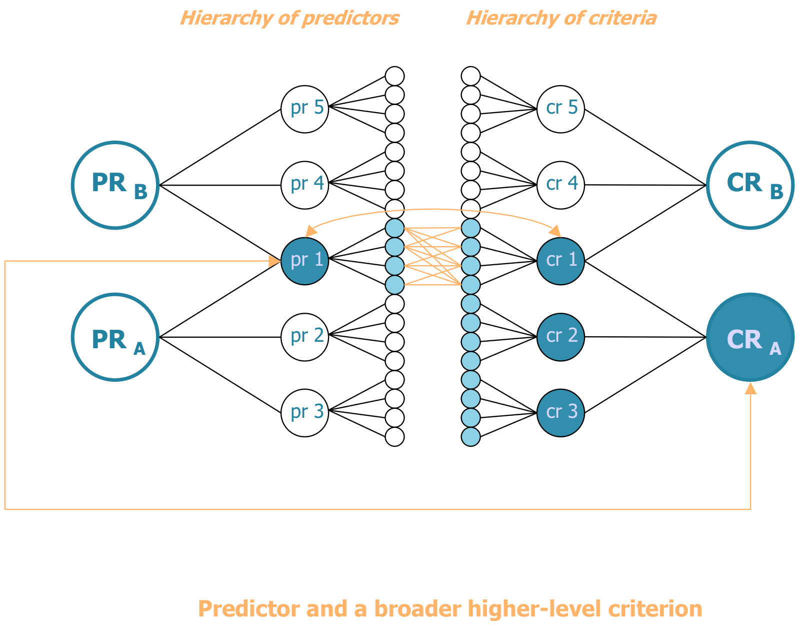
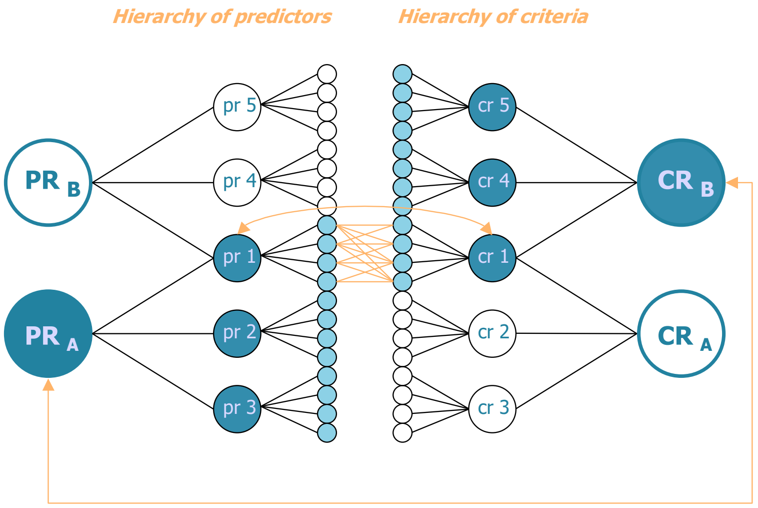


Figure 3d.

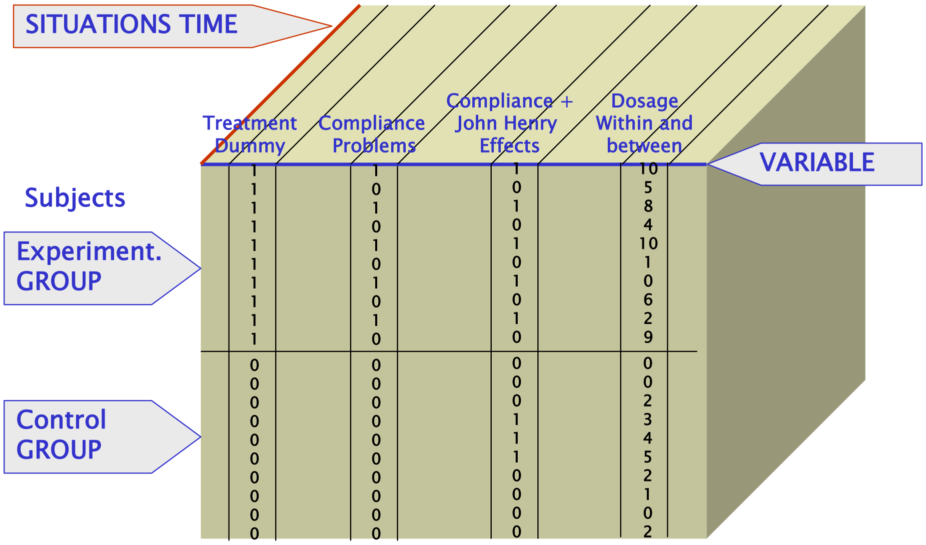
The hybrid case of asymmetry



Mismatch at the same level of generality !

Figure 4.

A closer look at the experimental treatment box



Treatment Dummy	Compliance Problems	Compliance + John Henry Effects	Dosage Within and between
1	1	1	10
1	0	0	5
1	1	1	8
1	0	0	4
1	1	1	10
1	0	0	1
1	1	1	0
1	0	0	6
1	1	1	2
1	0	0	9
0	0	0	0
0	0	0	0
0	0	0	2
0	0	1	3
0	0	1	4
0	0	1	5
0	0	0	2
0	0	0	1
0	0	0	0
0	0	0	2

Figure 5.

Nomogram for selection effects: Parameter S

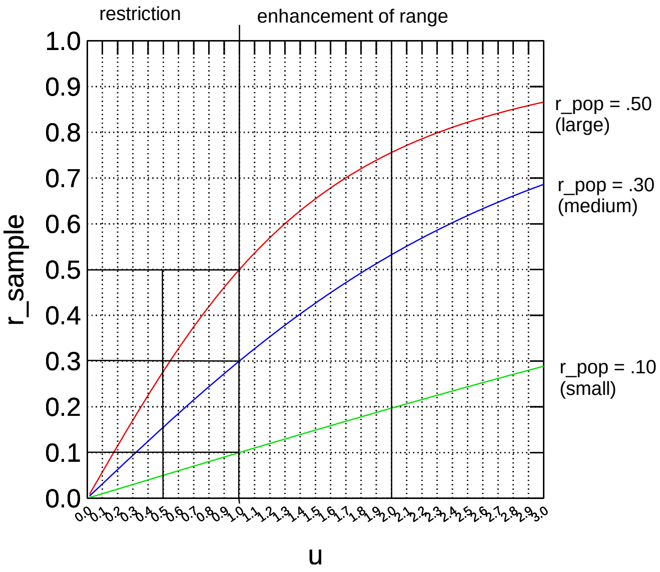
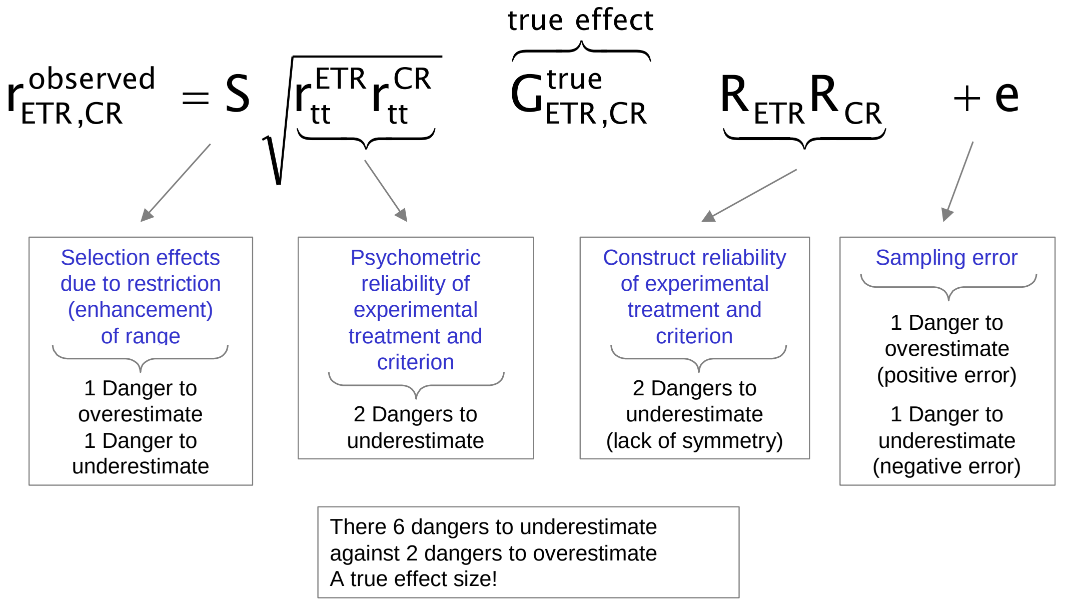


Figure 6.

The Brunswik-lens-equation for relating experimental treatment (ETR) to criteria (CR)

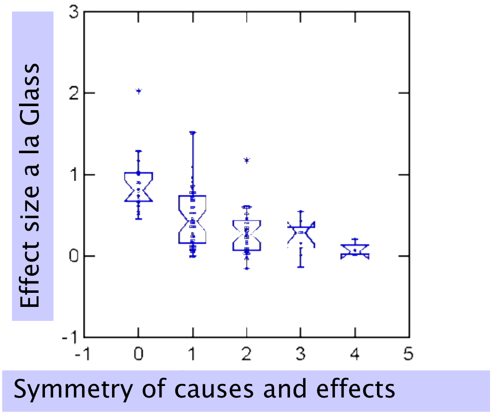


$$r^{observed}_{ETR,CR} = S\sqrt{\underbrace{r^{ETR}_{tt}\, r^{CR}_{tt}}}\ \overbrace{G^{true}_{ETR,CR}}^{\text{true effect}}\ \underbrace{R_{ETR}R_{CR}} + e$$

Selection effects due to restriction (enhancement) of range
1 Danger to overestimate
1 Danger to underestimate
Psychometric reliability of experimental treatment and criterion
2 Dangers to underestimate
Construct reliability of experimental treatment and criterion
2 Dangers to underestimate (lack of symmetry)
Sampling error
1 Danger to overestimate (positive error)
1 Danger to underestimate (negative error)
There 6 dangers to underestimate against 2 dangers to overestimate A true effect size!

Figure 7.

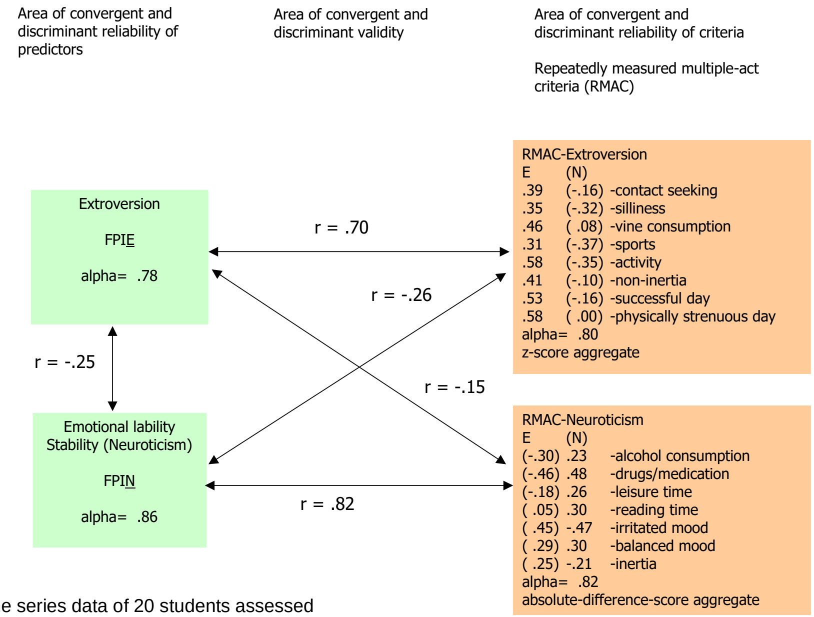
German psychotherapy effects as a function of symmetry



107 effects from Wittmann & Matt (1986) and the extention by Spinner (1991).
Low scores high symmetry.

Figure 8.

Testing Eysenck’s E-/N-theory in the Brunswik-symmetry framework[1]


Area of convergent and discriminant reliability of predictors
Area of convergent and discriminant validity
Area of convergent and discriminant reliability of criteria
Repeatedly measured multiple-act criteria (RMAC)
Extroversion
FPIE
alpha= .78
r = -.25
Emotional lability Stability (Neuroticism)
FPIN
alpha= .86
r = .70
r = -.26
r = -.15
r = .82
RMAC-Extroversion
E (N)
.39 (-.16) -contact seeking
.35 (-.32) -silliness
.46 (.08) -vine consumption
.31 (-.37) -sports
.58 (-.35) -activity
.41 (-.10) -non-inertia
.53 (-.16) -successful day
.58 (.00) -physically strenuous day
alpha= .80
z-score aggregate
RMAC-Neuroticism
E (N)
(-.30) .23 -alcohol consumption
(-.46) .48 -drugs/medication
(-.18) .26 -leisure time
(.05) .30 -reading time
(.45) -.47 -irritated mood
(.29) .30 -balanced mood
(.25) -.21 -inertia
alpha= .82
absolute-difference-score aggregate

[1] Time series data of 20 students assessed over 8 weeks from Fahrenberg et al. (1977)

Figure 9.

Behavior of Group Trainers as Perceived in Single Courses (plain lines) and on the Average (dotted lines)

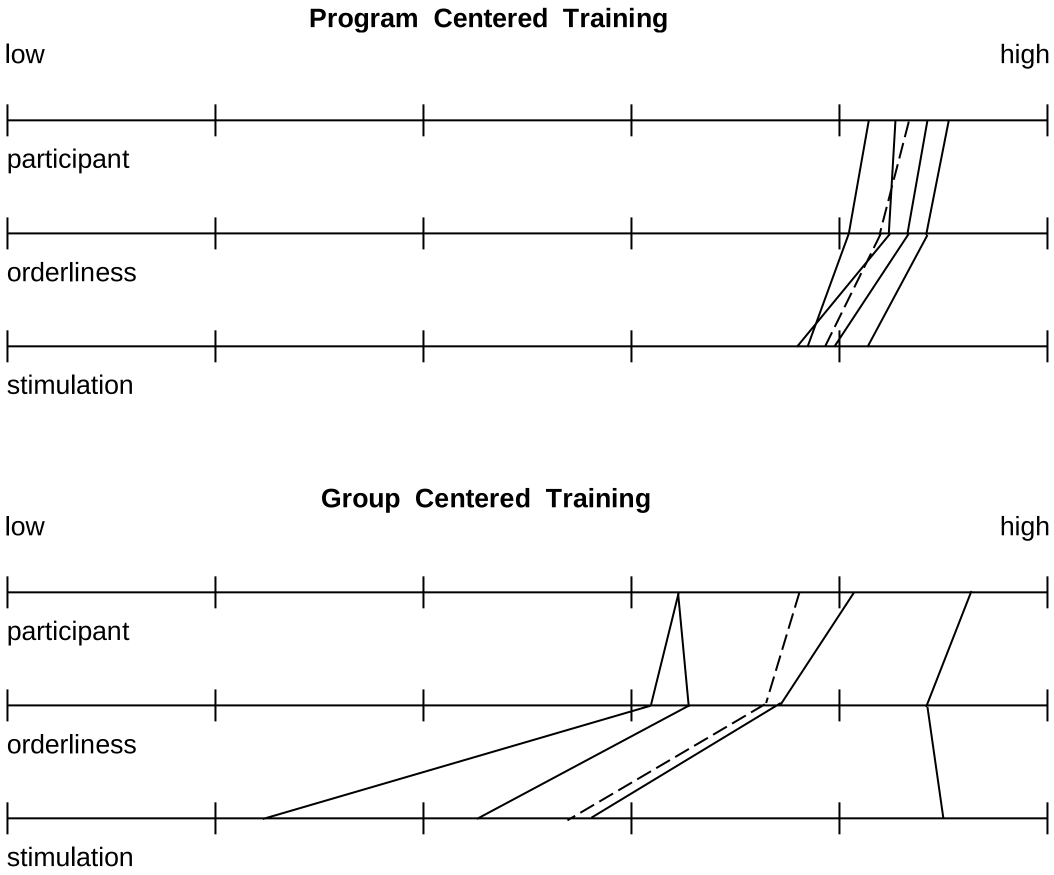
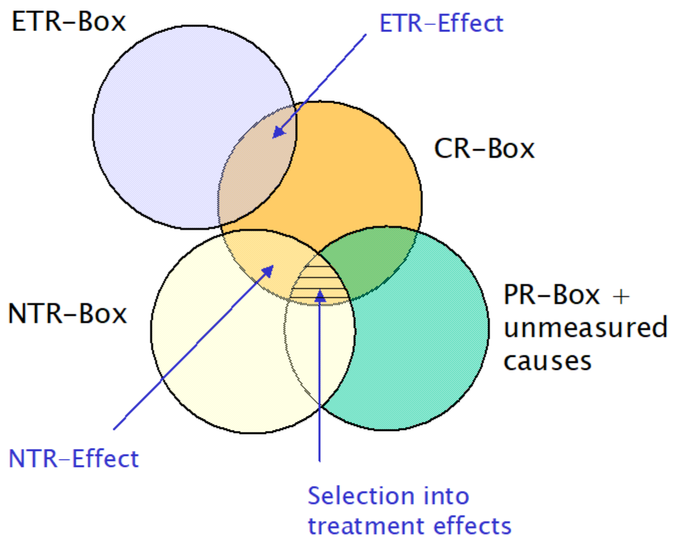


Figure 10.

Different effects using the five-data-box conceptualization


ETR–Box
ETR–Effect
CR–Box
NTR–Box
PR–Box +
unmeasured
causes
NTR–Effect
Selection into
treatment effects