A principle of derivation and decisions based on statistical tests and effect sizes

Willi Hager¹²

Abstract

This article in Chapter 1 deals with deriving predictions from one psychological hypothesis meant to be examined as aggregate hypothesis. During derivation, two additional criteria, adequacy and exhaustiveness, should be attended to. Adequacy refers to the direction of effects specified by the psychological hypothesis, and exhaustiveness to the number of partial predictions which the psychological hypothesis contains. Secondly, because of the lack of a logical form of judgement a strategy of decision and evaluation is outlined which allows to take into account more systematically measures of effect sizes in the evaluation of the psychological prediction and about the psychological hypothesis to be examined (Chapter 2). If an experiment is planned properly, there are only few patterns of results possible, where each pattern consists of the result of a statistical test and a value of the effect size. These possible patterns of results are tabulated, and each pattern is classified with respect to the meaning it bears for the psychological prediction derived from a psychological hypothesis. Under certain premises the evaluation of the psychological prediction can be transferred to the psychological hypothesis. – Whereas the derivation of predictions is largely based on deductive considerations, the evaluation of predictions and of the psychological hypothesis is a matter of inductive reasoning which never is as strict as deductive reasoning. Thus, the inductive reasoning part leaves more degrees of freedom for the researcher.

Key words: Derivation of predictions, effect sizes, decisions, evaluations

¹ Authors address: Prof. Dr. Willi Hager, Georg-August-Universität Göttingen, Georg-Elias-Müller-Institut für Psychologie, Goßlerstraße 14, D-37073 Goettingen, Germany, e-mail: whager@gwdg.de
² I thank E. Heise for her many helpful comments, and Shirley McCarthy for improving my English. The remaining mistakes are mine.
1. **A principle of derivation**

Although statistical tests have been misused in many cases, they will remain in use. Most of these (mis)applications consist in taking statistical hypotheses as tested by statistical test as the most important feature of an experiment. In my opinion the psychological hypotheses preceding all statistical hypotheses are the most important feature in a psychological experiment. Psychological hypotheses are tested by means of predictions derived from them. Before this background this article deals with how predictions testable with statistical tests should be derived from psychological hypotheses and how the empirical results should be used for evaluating psychological predictions and hypotheses. One central idea consists in the conviction that „deriving“ predictions is more or less a matter of deduction in the sense of Popper’s methodology, although the part logic plays may be smaller than expected – according to the analyses by Kuhn (1970) or by Westermann (1987); at least, even this „deduction“ rests on some necessary inductive premises (Westermann & Gerjets, 1994). The evaluation of the predictions and of the psychological hypothesis in part 2 is more like the „inductive reasoning“ preferred by Carnap (Stegmüller, 1973; Westermann, 1987, p. 112; Westermann & Gerjets, 1994) and cannot be performed with the same rigour as the derivation. This means that conclusions about predictions and hypotheses are necessarily less strict than deriving predictions.

1.1. **Psychological hypotheses referring to aggregates**

Chow (1991b, p. 1088) used the (psychological) hypothesis for illustrating his arguments „that information in the short-term store is acoustic in nature.“ He formulated this hypothesis without any reference to particular persons, situations, and so on. One might conclude that he considered this hypothesis to hold for all people or individuals and situations. Many psychological hypotheses refer to individuals or each individual of a particular class – they are „deterministic universal statements“ (Erdfelder & Bredenkamp, 1994; Westmeyer, 1998). But this may not be the way these hypotheses are understood – either by their „inventors“ or when they are examined (Westermann, 1987, pp. 42-43; 1988). Usually, one tolerates that there are some individuals for whom the hypothesis does not hold without calling the hypothesis „not valid.“ Thus, in research practice psychological hypotheses usually seem to be interpreted as „aggregate hypotheses,“ and Danziger (1987, 1990) shows the historical development of this re-
interpretation (see also Stam & Pasay, 1998). This probabilization can formally be introduced by supplementing the psychological hypotheses with a probability \( \pi \) (Westermann, 1987, pp. 42-43) or it can be introduced implicitly by examining the hypothesis by statistical tests which are based on probabilistic arguments and which usually refer to some kind of average or mean values.

The main reason for the probabilization lies in the fact that experimental validity, especially internal validity referring to the so-called *ceteris paribus* clause connected or to be connected with every psychological hypothesis, can best be approximated if the individuals have been randomized and if the data is averaged over the individuals (see Erdfeder & Bredenkamp, 1994; Westermann, 1987). The *ceteris paribus* clause claims „all other things being equal except those deliberately varied with respect to the hypothesis to be examined“ (see, for example, Meehl, 1991, pp. 16-18; Chow, 1991a, p. 349; Westermann, 1987, 1988); Steyer (1994, p. 670) speaks of *ceteris paribus distributionibus* meaning the the distributions of potential nuisance variables should be the same accross the experimental conditions. The goal of randomization as *the* important experimental technique to secure the *ceteris paribus* conditions is the better approximated the greater the number of units who or which can be randomized. Given the usually small sample sizes in empirical research it is likely that there will be (unsystematic) deviations from the *ceteris paribus* conditions. One way to take them into account is by means of null hypothesis significance testing (NHST) (Westermann, 1987, p. 120; see also Mayo, 1983, 1985) which can be looked upon as a tool to divide the whole of variation in the experiment into two parts, one called „unsystematic“ or „chance variation,“ and the other „systematic variation“ attributed to the variation of the experimental conditions.

The attempt to approximate internal validity leads to „weak“ examinations of a psychological hypothesis with respect to its claim to be valid for each individual of a class (where „weak“ is used as the opposite to the „severe“ tests addressed by Popper, 1992). Focusing on this claim, these hypotheses can be examined for individuals, using single case designs (cf. Kern, 1997). These scrutinies may be rather „strong“ with respect to the claim concerning the individuals, but rather „weak“ with respect to experimental validity. Both ways of examining have their strengths and their weaknesses, and the best way may be to perform both kinds of examination in succession as Westermann (1987, pp. 118-119) proposes. A particular psychological hypothesis can then be said to be examined for a hypothetical „average individual“ or for each individual of a given sample.
The psychological aggregate hypothesis may say that usually the information in short-term store or for most individuals is acoustic in nature, thus admitting that there may be some people for whom this statement does not hold. A possible prediction derivable from this hypothesis might claim that items presented acoustically are better remembered than items which have to be read. The statistical aggregate hypothesis then may refer to the mean number of items correctly remembered and says that on average the postulated relationship holds, that is, if the data is averaged over the individuals. Since relationships for individuals are not necessarily the same as those for an hypothetical „average individual“ (see Erdfelder & Bredenkamp, 1994), the examination of aggregate hypotheses „corresponding“ to the psychological hypothesis may lead to false evaluations of the psychological hypothesis. Nonetheless, it is customary (see Falk & Greenbaum 1995; Gigerenzer, 1993) to examine psychological hypotheses by testing statistical hypotheses referring to aggregates. I shall confine my further considerations to psychological hypotheses which are referred to an hypothetical average individual, i.e. I shall not consider predictions for single cases (see Hager, 1992a) nor other kinds of psychological hypotheses.

Testing psychological hypotheses by means of NHST then bears two outstanding advantages: Firstly, the statistical test allows to take into account the inevitable deviations from the ceteris paribus clause. Secondly, the Neyman-Pearson form of NHST (NPT) allows to control the probabilities $\alpha$ and $\beta$ for wrong decisions on the statistical hypotheses used to examine the psychological hypothesis.

1.2. The principle

There is wide agreement that hypotheses are examined by prediction derived from them (Carnap, 1946; Lakatos, 1970; Popper, 1992). According to Popper (1992, p. 32), these predictions are derived „by means of logical deduction,“ and they are compared with the results of an empirical experiment aimed at examining them. In addition, Popper (1992, pp. 75-76) states that these predictions are hypotheses with a lower degree of „universality“ (what Westmeyer, 1998, calls „Allgemeinheitsgrad“) than the hypothesis to be examined, and they are „singular statements“ (Popper, 1992, p. 60) referring to one particular situation or experiment. Empirical data that disagree with the predictions can serve to „falsify“ the hypothesis. Popper (1992, p. 86) also states:

A theory is to be called „empirical“ or „falsifiable“ if it divides the class of all possible basic statements unambiguously into the following two non-empty subclasses. First, the
class of all those basic statements with which it is inconsistent (or which it ... prohibits): we call this the class of the potential falsifiers of the theory ...; and secondly, the class of those basic statements which it does not contradict (or which it 'permits'). (emphasis in original)

The statistician R.A. Fisher, whose epistemological orientation was towards Carnap rather than Popper, has demanded that

the possible results of the experiment ... are divided into two classes with opposed interpretations. ... The two classes of results which are distinguished by our test of significance are ... those which show a significant discrepancy from a certain hypothesis ... and results which show no significant discrepancy from this hypothesis. (Fisher, 1966, pp. 15-16)

Although both scientists seem to have never cited each other, they share a common view, even if they may address different kinds of hypotheses. This fact gives rise to a further distinction concerning levels of prediction. A psychological hypothesis (PH) refers to non-observable constructs of psychology and to no experimental situation in particular. A psychological prediction (PP) is derived (\( \therefore \)) from the psychological hypothesis and should be constructed to differ from the psychological hypothesis in (at least) two respects: It refers to observable (psychological) variables such as a word list consisting of „high“ and „low imagery“ words (instead of the non-observable „imagery“) or an observable behavioral measure of (non-observable) aggression (such as „number of aggressive acts against other children during play“) and so on. In addition, a psychological prediction refers to an exactly specifiable experimental situation including a design layout, specific experimental manipulations and variables, subjects, and so on – in short: to a completely specifiable empirical system. As a consequence of this, the same psychological hypothesis will not necessarily lead to the same psychological prediction if at least one of the aspects addressed above is altered. But this fact does not affect the subsequent considerations. As is the case with most psychological hypotheses, a psychological prediction refers to the hypothetical average individual or to an aggregate, resp.

Usually, a psychological prediction specifies a whole pattern of results which completely agrees with a psychological hypothesis; in rare cases the pattern reduces to a single result. This pattern may be conceived of as a number of partial predictions the whole of which makes up the psychological prediction. This pattern of results is then „translated“ into one statistical prediction (SP), possibly consisting of various statistical partial hypotheses. When formulating the statistical prediction one disregards of the question
whether these statistical partial hypotheses are directly testable by a known statistical test or not. Statistical predictions quite often have to be decomposed into several testable statistical hypotheses (see below). If the prediction specifies a pattern of results, it may not be possible to find an \( H_0 \) or an \( H_1 \) which "corresponds" exactly to this pattern. Nonetheless, it is usually possible to formulate statistical hypotheses according to this pattern, even if they are not directly testable. These type of (partial) prediction then constitutes a fourth level which I call the level of statistical tests (ST). Given a certain psychological hypothesis and particular circumstances of its testing, it is always one or more \( H_0 \) or \( H_1 \) of a known statistical test which serves as a prediction. In some instances, however, this statistical prediction is equivalent to a testable statistical hypothesis, i.e., either to an \( H_0 \) or an \( H_1 \). No decomposition is necessary then. – On the whole, there is one psychological hypothesis, one psychological prediction, eventually consisting of some partial predictions, one statistical prediction, which may also consist of several partial predictions, and there are one more testable \( H_0 \) and/or \( H_1 \).

Calling certain statistical hypotheses predictions means that they are intended to serve the function Popper (1992, p. 100) has assigned to "basic sentences," which are the basis for evaluating (psychological) hypotheses and theories or which can be used as "possible falsifiers" (see above). In order to serve this function, however, both kinds of statistical hypotheses must be given equal "importance," although the statistical tests always focus on "their" \( H_0 \) (see Hager, 2000).

Predictions can be those states that completely agree with the psychological hypothesis or that completely disagree with it. In either case the class of all possible empirical results is divided into mutually exclusive sets. The greater the class of results disagreeing with the hypothesis, i.e. results or pattern of results it prohibits, the greater its empirical content according to Popper (1992, pp. 119-121). In order to appropriately scrutinize a psychological hypothesis it is necessary to conserve its meaning or empirical content during derivation of predictions as far as this is possible when substituting non-observable variables or constructs by their observable counterparts, i.e. by their operationalizations. But even if a researcher considers this demand, he or she has more degrees of freedom than seems to be generally realized to choose a particular empirical system to which the theory is to be applied. In spite of these degrees of freedom I shall restrict my considerations to some basic cases and consider the agreeing predictions as is usually done.
In order to ensure that the empirical content is conserved during derivations or that „hypothesis validity“ in the sense given by Wampold, Davis and Good. (1990) is guaranteed, a „principle of derivation“ is formulated, where „derivation“ is understood in „a rather loose sense“ after Meehl (1967). This principle may be stated as follows: „Always choose your predictions with respect to the psychological hypothesis and to the empirical system or to the particular side conditions adequately, exhaustively, and in a way that ensures unambiguous division of all possible results.“ The principle is meant to preserve as much empirical content of the psychological hypothesis as possible given the inevitable loss of content when deriving predictions from a psychological hypothesis. It does not deal with „appropriately“ choosing the statistical concepts (such as parameters or distribution functions), but only with how to formulate relationships between the parameters or distribution functions according to what the psychological hypothesis says. This choice must depend on considerations outside the scope of this article, for example, on whether one takes variances or correlations or expected means or the like as entities to relate, which – in turn – may depend on what the PH actually asserts and what scale level is necessary for particular assertions to be empirically meaningful and so on. I refer to the hypothetical average individual and to means or average values, interpreted as the expectation of a random variable which may conventionally be considered as being normally distributed over all possible repetitions of the experiment.

Besides the demand of dividing the possible empirical results unambiguously into two sets with opposite meanings for the psychological hypothesis, the principle encompasses two central aspects, namely „adequacy“ and „exhaustiveness“ of derivation or formulation of predictions. The criterion of „adequacy“ refers to whether the psychological hypothesis claims a relationship or the lack of a relationship between the variables under consideration. It also refers to whether this claim is directional („something changes in a certain direction“) or non-directional („something changes“ or „nothing changes“). Adequacy is violated when testing a directional psychological hypothesis by a nondirectional statistical hypothesis.

The second criterion is called „exhaustiveness.“ This means that the number of relationships addressed by the psychological hypothesis must show up in the predictions. If a quantitative psychological hypothesis predicts a quantitative linear trend for some experimental conditions, and if this prediction is tested by means of a test for a (qualitative) monotonous trend, the (implicit) derivation is not exhaustive. Linearity of trend means that the mean values of the dependent variable can be monotonously ordered according to the values of the independent variable, and that the distances between any
two mean values of the dependent variable are proportional to the respective differences of the independent variable. A monotonous trend, on the other hand, is defined as a strict or weak ordering of the mean values of the dependent variable, but does not allow any statement to be made about the distances between these mean values (see Hager, 1996, for the details). If – as is very often the case – the psychological hypothesis leads to the prediction of a monotonous trend for – say – three experimental conditions the application of a test for quantitative trends is also not exhaustive, but „in the other direction:“ This test (implicitly) enhances the empirical content of the psychological hypothesis in that it is sensitive to distances between means values whereas the hypothesis does not say anything about these distances. Most often, however, the criterion of exhaustiveness is violated, when dealing with interaction hypotheses (see below).

1.3. Formulation of predictions

Before being able to consider possible predictions, one has to decide which techniques of experimental control one can apply to enhance experimental validity (Bredenkamp, 1980; Cook & Campbell, 1979; Gadenne, 1976; Hager, 1998; Westermann, 1987). If a potential nuisance variable is known one may vary it systematically and introduce it into the design as a control factor. This enables more predictions to be tested than when considering only the variables which are necessary with respect to the psychological hypotheses, while the known nuisance variable is held constant over the experimental conditions. All aspects of controlling for nuisance variables are part of the empirical system for which the PH is examined, and these aspects affect the design layout very often (design with only one factor vs. design with one factor referring to the hypothesis and a second factor introduced for control, and so on). Thus very different design layouts may be used to examine the same psychological hypothesis, and each of these designs forms another empirical system. I will not address more complex cases, but rather restrict my considerations to some basic cases (see Hager, 1992a, for further details).

In order to demonstrate the derivation with a very simple example, let us consider the psychological hypothesis „The higher the imagery content of material to be learned the higher the amount of learning“ (Paivio, 1986). The respective theory specifies in which way the construct „imagery“ may be operationalized and how to assess „amount of learning“ empirically. Subsequently a way of presenting the word lists is chosen, an instruction, subjects, randomization, a one-way layout with two conditions („high“ vs. „low imagery“, as assessed by the mean imagery values of the word lists), and so on. Let us assume that the operationalization of „amount of learning“ is given by a respective
learning test, asking the subjects to recollect as many words learned as possible (I disregard the distinction made between different learning processes which lead to different ways of how to assess learning). For this operationalization and design layout and so on, the psychological prediction may be: \((PH \land AH) \Rightarrow PP: \text{"The number of words correctly recalled is greater with for a word list with mean imagery ranking of } 5.7 \text{ (high imagery) than in a wordlist with mean imagery ranking } 2.5 \text{ (low imagery)."}"

This prediction does not address single subjects or a hypothetical average subject. If the first interpretation is chosen, each single subject should learn high imagery words as well as low imagery words (within-subject variation), and it is predicted that the number of words correctly recalled is higher for high imagery words than for low imagery words – for each subject. If one tolerates that some subjects do not behave according to the prediction – maybe because of violations of the ceteris paribus clause –, one can perform a binomial test, thus introducing probabilistic criteria. If one would not tolerate deviations no test seems necessary. One might, however, define a minimum difference between high and low imagery words recalled correctly in order to take inevitable violations of the ceteris paribus clause into account. A statistical test is also not necessary in this case. – The second interpretation leads to predicting that the mean number of words correctly recalled is higher for high imagery words (mean ranking 5.7) than for low imagery words (mean ranking 2.5) ("aggregate hypothesis"), which means that two hypothetical average individuals are compared (between-subjects variation) or that the average number of words correctly recalled with the two lists is compared for an average individual (within-subject variation). Although in most cases the type of variation of the experimental conditions (within subjects vs. between subjects) has no effect on deriving the predictions, but merely constitutes a particular procedure being part of the empirical system examined (which means that results may differ for both types of variation!), there are other cases when it does; an instructive example for this is given by Bredenkamp (1980, pp. 48-49).

Under the second interpretation addressed above, the scores in the learning test are interpreted as experimental realizations of a random variable with a particular distribution often assumed to be normal. The statistical prediction then refers to the expected mean of this random variable \((\mu)\), but quite often the observable dependent variable is not of the scale type necessary to lead to empirically meaningful statements (in the sense of measurement theory) about the expected means. In cases like these one may consider operationalizations of "average" different from the preceding one, e.g. use medians instead of expected means. If one uses a statistical hypothesis referring to medians or
mean ranks as the aggregate hypothesis, it might occur in some rare instances that because of the median paradox identified by Iseler (1996) no proper testing of the psychological hypothesis seems possible. Taking into account the (present) impossibility of controlling $\beta$ for rank tests, it seems preferable to use a statistical hypothesis as a prediction which can be tested by a binomial tests or the like whose power can easily be determined and computed. On the other hand, the researcher may conclude that the median paradox is unlikely to occur to him or her under the circumstances given so that she or he can consider medians as the statistical constructs of interest; the psychological hypothesis is then examined by means of a statistical hypothesis about medians. Of course, other options or auxiliary hypotheses are possible, but will not be considered here. – Let us return to the statistical prediction (SP) derived above.

It says that $\mu_{III}$ will be higher than $\mu_{II}$ or: $\mu_{III} - \mu_{II} > 0$ and it is equivalent to the (directional) $H_1$ of a (one-sided) $t$ test.\(^3\) The psychological hypothesis (PH-A) and some appropriate auxiliary hypotheses (AH and AH') subsequently to be omitted, however, lead to PP-A and to SP-A, which is equal to the assertion that a particular $H_{I,A}$ must hold if PH-A is true (for the given experiment or empirical system) (see also Meehl, 1967):

\[(PH-A \land AH) \Rightarrow (PP-A \land AH') \Rightarrow SP-A: (\mu_{III} - \mu_{II} > 0) \Leftrightarrow H_{I,A}\]

or for short

\[PH-A \Rightarrow PP-A \Rightarrow SP-A: (\mu_{III} - \mu_{II} > 0) \Leftrightarrow H_{I,A}\]

("$\Rightarrow$" here and subsequently denotes the loose derivation of conforming predictions or patterns of results, without explicit reference to the necessary auxiliary hypotheses AH and AH' in the second line; "$\Leftrightarrow$" means "is equivalent or identical to"). More "loosely" stated one may say that the PH-A is examined by means of particular predictions including the particular $H_{I,A}$, and this means that the PH-A may be examined by other predictions as well. Therefore, the particular predictions constitute a part of the empiri-

\(^3\) Sometimes, it is argued that one-sided tests (on directional statistical hypotheses) are problematic since it may be that there is a "significant difference" in the wrong direction. Whether the difference in the "wrong" direction be statistically significant or not, it always disagrees with the predictions, and there seems no reason to discuss its statistical significance with respect to its meaning concerning the psychological hypothesis under scrutiny. But it may be possible to take this deviation into the "wrong" direction (whether it is significant or not) as a basis for trying to explain this deviation and/or to formulate a new hypothesis to be examined in a further investigation (but see Cowles, 1989, pp. 178-180).
cal system which the PH-A refers to. With only two experimental conditions the criterion of exhaustiveness cannot be violated. By formulating a directional statistical hypothesis the criterion of adequacy is fulfilled. Table 1 contains the predictions.

<table>
<thead>
<tr>
<th>Psychological hypothesis (level of PH)</th>
<th>High imagery words are learned better than low imagery words</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>high imagery(5,7) Prediction low imagery (2,5)</td>
</tr>
<tr>
<td>Psychological prediction (level of PP)</td>
<td>CRW_{HII} &gt; CRW_{LII}</td>
</tr>
<tr>
<td>Statistical prediction (level of SP)</td>
<td>\mu_{HII} &gt; \mu_{LII}</td>
</tr>
<tr>
<td>Statistical tests (level of ST)</td>
<td>H_0: \mu_{HII} &gt; \mu_{LII}</td>
</tr>
</tbody>
</table>

*Notes.* CRW means (mean) number of correctly recognized words. The levels of ST and SP need not be distinguished in the case of only two experimental conditions. But in case of more than two conditions this differentiation is necessary (see above).

Table 1: A four-step scheme for deriving predictions

In other instances, an H_0 is linked as conforming to a psychological hypothesis (PH-B) such as „High imagery leads to the same memory performance as high concreteness does:“

\[
\text{PH-B} \Rightarrow \text{PP-B} \Rightarrow \text{SP-B}: (\mu_{HII} - \mu_{HC} = 0) \Leftrightarrow H_{0,B}.
\]

This example shows, by the way, that the possibility of deriving null hypotheses from a psychological hypothesis or theory occurs more often even in the „soft“ areas of psychology than the articles by Meehl (1967, 1978, 1991) might suggest (see Brom, Defares & Kleber, 1989, for a case in point in therapy research).

Subsequently, I will not explicitly deal with the psychological predictions, but rather the statistical ones. Let us consider two more complex cases.

The PH-C again claims a relationship between a qualitative independent variable (such as „imagery“) and a dependent variable (such as „amount of learning“). The researcher decides to constitute three levels of the independent variable, operationalized as imagery content in three word lists so that the average imagery content of the word lists serves as a numerical value for „high“ (HI; 5.7), „medium“ (MI; 4.0), and „low imagery“ (LI; 2.5). She or he uses a recognition test as operationalization of the dependent variable and counts the number of correctly recollected words to which the statistical constructs (\(\mu_i\)) refer. One can construct the subsequent chain of derivation relative to the PH-C and the side conditions chosen:
PH-C \Rightarrow PP-C \Rightarrow
SP-C: (\mu_{III} > \mu_{III} > \mu_{LI}) \Rightarrow [(H_{I;1}: \mu_{III} > 0) \wedge (H_{I;2}: \mu_{III} = \mu_{LI} > 0)],

where \(^{\wedge}\) signifies the logical conjunction between the statistical partial hypotheses. Whenever a decomposition of the statistical prediction into testable statistical partial hypotheses is performed it is necessary to define a decision rule or criterion which specifies the conditions when to accept or to reject the statistical prediction. The conjunctive linkage of the two testable alternative hypotheses serves as a severe decision criterion. It says: "Accept the statistical prediction only if all the statistical partial hypotheses it has been decomposed to can be accepted." In the case of a lenient or benevolent disjunctive linkage (\(^{\vee}\)) between the two partial hypotheses the decision criterion says: "Accept the statistical prediction if at least one of the partial hypotheses constituting it can be accepted." This leads to a greater benevolence of the test.

Let us consider a last example. The PH-D claims that "the effects of high imagery are especially pronounced for concrete words." This leads to the following statistical prediction SP-D:

PH-D \Rightarrow PP-D \Rightarrow SP-D \Rightarrow
\{(H_{I;1}: \mu_{11} - \mu_{12} > 0) \wedge (H_{I;2}: \mu_{21} - \mu_{22} > 0) \wedge [H_{I;3}: (\mu_{21} - \mu_{22}) - (\mu_{12} - \mu_{11}) > 0]\}

These partial predictions are illustrated in Table 2.

<table>
<thead>
<tr>
<th>Concreteness</th>
<th>Imagery</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>high (5.7)</td>
</tr>
<tr>
<td>abstract</td>
<td>\mu_{11}</td>
</tr>
<tr>
<td>concrete</td>
<td>\mu_{21}</td>
</tr>
</tbody>
</table>

Table 2: Illustration of the statistical prediction SP-D

The two tests for the simple effects are necessary for establishing the imagery effect, and the interaction test is necessary to show that the imagery effect is particularly pronounced when the words are concrete. There are no predictions concerning the two main effects (Imagery and Concreteness), although a directional main effect for imagery can be deduced. But the corresponding main effects test need not be performed, since the two tests for the simple effects grasp the relevant information. On the other hand, omitting these tests will lead to a non-exhaustive derivation, because the PH refers to an imagery effect which is expected for both conditions of the concreteness factor. It may
happen that the interaction test comes up statistically significant, although one simple
effect or both of them is counter to predictions.

1.4. Effect sizes

The outcome of any statistical test depends on the sample size $N$, the level of signifi-
cance $\alpha$, the power $1 - \beta$, and the actual effect size (ES). There seems to be no psycho-
logical hypothesis which gives us any information about the sample size „appropriate for
its testing,“ and/or about the „appropriate“ level of significance, or the „appropriate“
power. These values, then, reflect specifications that have to be made outside the psy-
chological theory or hypothesis on methodical or „technical“ grounds and that depend
on the resources of the experimenter as well as on the customs in the particular area of
research (cf. Kraemer, 1998). But psychological hypotheses tell us whether experimental
manipulations should lead to a difference between the experimental groups or to no dif-
fERENCE, i.e., they tell us whether a substantial effect is to be expected ($\text{PH} \Rightarrow \text{PP} \Rightarrow \text{SP} \Leftrightarrow H_1$) or whether no (substantial) effect is to be expected ($\text{PH} \Rightarrow \text{PP} \Rightarrow \text{SP} \Leftrightarrow H_0$).
This means that of all determinants of statistical tests the effect size can be viewed as
the statistical concept that shows closest relationship to the psychological concepts. This
is not to say that the relationship is close, but it is closer than is the case with the other
determinants of a statistical test. NHST tells us whether there is a statistical effect in
the data or not, and the measure of effect size tells us how big this effect is.

Thus, statistical tests and effect sizes give us complementary, although partly over-
lapping, pieces of information: Whereas the statistical test depends heavily on probabili-
ties and on sample sizes, the effect size is mostly free of such aspects (though sampling
distributions can be analytically derived for many effect sizes). Therefore, effect sizes
and statistical tests are no alternatives, as Chow (1988, 1991a) claims, and effect sizes
should not be disposed of, as he proposes, but should be incorporated in evaluations more
systematically than is the case nowadays: „Scientific research workers ... pay undue at-
tention to the results of tests of significance ... and too little to the estimates of the ma-
gnitude of the effect they are investigating“ (F. Yates, 1951, p. 31).

On the other hand, no qualitative psychological hypothesis I know of allows derivati-
on of a certain value of the effect size as a prediction. When considering a quantitative
psychological hypothesis, however, it is usually possible to determine exactly the statis-
tical effect to be expected when the hypothesis holds true. This follows from the fact
that quantitative hypotheses refer to functional relationships which not only enable pre-
dicting (mean) values of the dependent variable or certain correlations or the like, but also exact values of effect sizes in the form of, say, correlations (Bredenkamp, 1972). Consider a quantitative hypothesis about a positive linear trend, which allows predicting a positive linear correlation ($\rho$) of an exact size, depending on the values of the independent variable and the variance within the experimental conditions. One possibility consists in testing the exact prediction by means of $H_0: \rho = c$. To plan this test, one has to fix a narrow range of deviations from this prediction which one thinks small enough to be in accordance with the prediction (Serlin & Lapsley, 1993). These deviations are fixed as another effect size not being derivable from the theory or hypothesis. Therefore, even in the case of quantitative psychological hypotheses examined by means of statistical ones the size of at least one effect has to be specified on grounds outside the hypothesis, that is on methodical grounds as a "technical" effect – despite some assertions to the contrary (e.g., by Cohen, 1988).

As is the case with any $H_0$ derived as a prediction the only value fully in agreement with this prediction is zero (or another exact value like $c$ used precedingly), since the $H_0$ usually asserts that no effect is to be expected, that two categories are (stochastically) independent, that a relation is absent, and so on. But since power $1 - \beta$ cannot be determined for any zero effect or zero correlation, a region encompassing small values of effect sizes has to be specified which do not fully agree with the exact prediction, but which are small enough to be considered not to contradict the exact prediction of a zero effect (see Hager, 1992a, 1993; Serlin & Lapsley, 1993). This means that the effect size which the test is planned to detect should be small if the psychological hypothesis is examined via a statistical $H_0$; the effect size has to be interpreted as maximum tolerable deviations from the exact $H_0$. It should be greater if the psychological hypothesis is tested by means of an $H_1$; here, effect sizes have to be interpreted as minimum effects just large enough to be of substantive importance.

But since planning of tests usually means finding a compromise for several demands not all of which seem compatible if faced with limited resources, it does not seem possible to give short-cut rules which concerning the magnitudes of effects: When is it "small" or "large" in general? In some areas researchers might find it appropriate to detect "small" effects when testing a psychological hypothesis by means of a derived $H_1$ that researchers in another area think appropriate to relate to their psychological hypothesis which is scrutinized by means of an $H_0$. The decision concerning "small" and "large" values has to be made individually (Newcomer, 1994, p. 394-395) and considered
separately for each dependent variable – some recommendations for this choice are given by Hager (1995, pp. 183-188) and by Howell (1997).

Moreover, it must be kept in mind that all too often the effect sizes cannot be chosen freely if the researcher thinks it necessary to keep her or his \( \alpha \) values within conventional limits and if she or he is faced with limited sample sizes, as seems to be usual in psychology. If the size of the sample and \( \alpha \) are fixed one can only vary the values of \( \beta \) and of the effect size. If one accepts Cohen’s criterion that \( \beta \) should not exceed the value .20 (Cohen, 1988), the detectable effect is fixed, too, since it can be computed for any given specifications of \( \alpha \), \( \beta \), and \( n \) – at least as far as many parametric and some non-parametric tests are concerned. This in turn leads to the recommendation to prefer statistical hypotheses as predictions which can be tested by statistical tests which can be planned accordingly, if this is possible and if there are different choices.

As has been argued, in order to plan tests or to analyze power, values of effect sizes should always be considered. Since values of effect sizes cannot be derived from the psychological hypotheses, specifications according to these values usually are more or less arbitrary, especially as they have to be made prior to experimentation. This arbitrariness seems to prevent many researchers from systematically dealing with effect sizes prior to experimentation (see Howell’s, 1997, p. 218, comments on this point). Even to deal with arbitrary specifications prior to actual experimentation seems necessary on the one hand because refraining from explicit specifications in advance only means that they implicitly take place by choosing a certain sample size and level of significance (see above). What I argue for is that these values be made explicit so that they can be discussed by others – an idea cogently formulated and justified by Bredenkamp (1972). On the other hand, the very arbitrariness of some (if not all) specifications of decision criteria in advance is in complete accordance with the opinion held by Popper (1981, p. 38):

But real support can be obtained only from observations undertaken as tests (by ‘attempted refutations’); and for this purpose criteria of refutation have to be laid down beforehand: It must be agreed upon which observable situations, if actually observed means that the theory is refuted. (emphasis in original)

And this means that „... it is decisions which settle the fate of theories“ (Popper, 1992, p. 108; emphasis in original). This encompasses the case where the detectable effect is merely a consequence of a fixed sample size, of a fixed level of significance, and of fixed power – a „technical“ effect (see above).
Overall, statistical tests should be planned for the specifications addressed (error probabilities, "technical" and non-technical effect sizes, and sample sizes; see Cohen, 1990, p. 1310).

2. Decisions and evaluations

The predictions include psychological (PP) and statistical ones (SP) and testable statistical hypotheses (ST) so that four levels can be constructed, beginning with the testable statistical hypotheses, statistical prediction, psychological prediction, and psychological hypothesis (see Table 3). Looked upon from a decision theoretic point of view, the decision on the statistical hypotheses is a "choice ... between well-defined sets of alternatives," whereas any statements concerning the psychological prediction is an "evaluation ... of one alternative at a time" (J.F. Yates, 1990, p. 3).

The separation of various levels of judgements (see Table 3) seems necessary since some of the problems discussed, for example, by Chow (1988, 1989) and by Folger (1989), can be traced back to the fact that these authors try to incorporate information contained in effect sizes into the decision about statistical hypotheses, whereas in the strategy proposed here information about the effect sizes is incorporated outside statistical decisions – strategies of the former kind have been discussed by Westermann and Hager (1982).

| Decision on the statistical hypotheses (ST) | accept or reject |
| Decision on the statistical prediction (SP) | accept or reject |
| Evaluation of the psychological prediction (PP) | shows up or does not show up |
| Evaluation of the psychological hypothesis (PH) | corroborated or not corroborated |

Table 3: Decisions and evaluations on four levels of consideration

2.1. Decisions on the tested statistical hypotheses and on the statistical prediction

There is no widely accepted logical form for judgements or for conclusions which is as strict as derivation of predictions. The reason for this lies in the fact that an evaluation of a psychological prediction and hypothesis based on statistical hypotheses always encompasses some inductive arguments, since the construct addressed by the psychological hypothesis encompasses more than the operationalization addressed by the psychological
prediction (see Hager, 2000). So, there is some need to develop a strategy which at least enables some decisions and evaluations.

Provided the test has been planned to control for $\alpha$ and $\beta$ the decision on the hypotheses tested is made according to the well-known principle addressed above: „Reject $H_0$ and accept $H_1$, if the empirical value of your test statistic falls in the region of rejection and retain $H_0$ in all other cases.” The region of rejection has been defined by the statistical hypothesis of interest and by the test administered, and its size corresponds to $\alpha$. This decision concerns a particular basic statement used to examine a psychological hypothesis, and the acceptance or rejection of it is reached by means of conventionally accepted criteria.

If only one hypothesis serves as a prediction the foregoing decision is also the decision on acceptance or rejection for the statistical prediction. If the statistical prediction has been decomposed into several partial hypotheses, the decision on its rejection or acceptance rests on the outcomes of the statistical tests and on the decision rule defined with respect to the statistical prediction, i.e. the conjunctive or disjunctive linkage of the hypotheses actually tested (see above). If there are many such hypotheses and if the conjunctive criterion has been chosen, the researcher may well decide to tolerate a small number of tests not agreeing with the predictions without rejecting the statistical prediction. This means giving up the „strong“ criterion, but that may be justified by the argument that there will be a usually small number of wrong decisions on statistical hypotheses by mere chance. Moreover, the researcher may decide to accept the statistical prediction if all differences point in the predicted direction, although not all of them have been judged to be significant. She or he may even decide to accept the statistical prediction although there are a few results in the direction opposite to predictions. Other modifications of the decision rule are possible, but will not be considered here. One may say however, that the scrutiny of the psychological hypothesis may be the less severe the more the criterion of decision is undermined, but it may be more benevolent.

The acceptance of the statistical prediction makes it easier to justify a positive evaluation of the psychological hypothesis, whereas the rejection of the statistical prediction makes it easier to justify a negative evaluation of the psychological hypothesis. Neither result will be called a „necessary condition“ for a particular evaluation of the psychological hypothesis.
2.2. Evaluation of the psychological prediction

The decisions addressed up to now have been rather mechanistic and are based on criteria set in advance. Many authors claim that “informed reasoning” (Falk & Greenbaum, 1995, p. 94) and “thoughtful decisions” or “educated guesses” (Hays, 1988) and “informed judgments” (Cohen, 1990) (see above) should replace the mere ritualistic or mechanic decisions or evaluations based on tests of significance (see also, e.g., Gigerenzer, 1993; Howell, 1997), and pursuing this demand is essential for the subsequent evaluations which are based on the preceding decisions. As is argued in this paper, “informed reasoning” should also be applied when using NHST, even if this testing only results in a formal decision, but it should predominantly be applied to the evaluations concerning the statements following the decisions on statistical hypotheses.

The psychological prediction refers to psychological variables or observable entities. Thus deciding on the psychological prediction means interpreting the outcomes of the statistical tests with respect to psychological variables, and this step contains inductive aspects since the psychological variables contain more meaning than the statistical concepts. As has been argued, the effect size may be said to have the closest connection to psychological concepts. Thus, it makes sense to take it into account when evaluating the psychological prediction.

Accepting the statistical prediction in most cases tells us that there are certain statistical effects, that is, if PH-A ⇒ PP-A ⇒ SP-A ⇔ H_{1,A}. If PH-B ⇒ PP-B ⇒ SP-B ⇔ H_{0,B}, no effect of substantial size is expected, and accepting the H_{θ} of a test appropriately planned always means that the empirical effect size is smaller than the one defined as a criterion value for planning the test (Hager, 1987; Westermann & Hager, 1982). This, of course, lowers the actual power but the researcher may be confident that the effect is small enough to be in accordance with the H_{θ} retained. Thus, if PH-B ⇒ PP-B ⇒ SP-B ⇒ H_{0,B} and if this H_{θ} is retained, the respective SP-B and the PP-B may be said to hold. If PH-A ⇒ PP-A ⇒ SP-A ⇒ H_{1,A} very little is said about the size of the effect, given a wide range of sample sizes to be found in psychological experiments. Thus, the empirical effect size (ES_{emp}) has to be computed and should be compared with the criterion value (ES_{crit}) used when planning the test. If SP ⇔ H_{1}, that is, if no decomposition is necessary, the empirical effect can be smaller or greater than the criterion value. If the latter is the case, the psychological prediction can be said to agree with the data. If the former is the case, the evaluation of the psychological prediction may depend on the magnitude of the difference between ES_{emp} and ES_{crit}. If this difference is
small there should be a positive evaluation in most cases; if it is of considerable size you may decide that it is too small to be of interest with respect to the psychological prediction. But despite this large difference one may call the psychological prediction to have shown up if one thinks it to be justifiable. These proposals and others below are dependent on the effect sizes chosen to be suited to the test statistic. So, the standardized mean difference, $\delta$, is suited for the t test, and the squared multiple correlation, $\eta^2$, is suited for the F-test and so on.

Another detail should be considered to. If PH-A $\Rightarrow$ PP-A $\Rightarrow$ SP-A $\Leftrightarrow$ $H_{1,A}$, the larger the effect, specified in advance, the more unlikely it is that there will be a positive evaluation of the PP-A, if one compares the empirical effect to the criterion value as recommended, since comparatively small effects are the rule rather than the exception in psychological research. But large effects can be detected with greater power than small effects – all other things being equal, whereas greater sample sizes are needed to detect smaller effects. Therefore, lowering the criterion effect size enlarges necessary sample size, but enhances the probability of a positive evaluation of the psychological prediction. If PH-B $\Rightarrow$ PP-B $\Rightarrow$ SP-B $\Leftrightarrow$ $H_{0,B}$, usually small „technical“ effects are focused on as a belt of tolerable maximum deviations from the $H_{0,B}$. The power is determined for the single exact maximum deviation, resulting in large sample sizes. In order to reduce it the maximum deviations may be enlarged, but this possibly means defining a region of deviations which may be said to be too large with respect to the prediction of a lack of difference or the like. There is no simple rule for coping with these difficulties, but what is necessary again is an explicit decision of the researcher to fix her or his criteria with respect to the manageable sample size. Overall, greater resources are needed if one wants to examine a psychological hypothesis by means of an $H_0$ than by means of an $H_1$, and often it is wise to consider different design layouts or ways of examining, if faced with limited resources.

In addition, the preceding considerations indicate that it may not make much sense to associate „practical (or clinical or ...) significance“ or „importance“ with the sizes of effects (e.g., Kirk, 1996). A small or minor effect may be of great importance if PH-B $\Rightarrow$ PP-B $\Rightarrow$ SP-B $\Leftrightarrow$ $H_{0,B}$ as is often the case when comparing, e.g., the comparative effectiveness of programs (such as therapies) with the same goals (Hager, 1999). On the other hand, larger effects may be of importance if PH-A $\Rightarrow$ PP-A $\Rightarrow$ SP-A $\Leftrightarrow$ $H_{1,A}$, but even in this case it often makes sense to try to detect small effects, if, for example, the phenomenon to be examined is difficult to evoke or is elusive or the like (see also Prentice & Miller, 1992). Thus, the „significance“ (or importance) of effects can only be judged
in relationship to a particular psychological hypothesis and the way it is actually tested, including the operationalizations used. There is no clear-cut rule as to which effects are important or not – their importance is only loosely linked to their magnitudes.

Assuming that the data has lead to accepting the $H_1$ and that the test has been planned to be able to detect a standardized mean difference $\delta_{\text{crit}} = 1.0$, a rather large effect. The empirical effect $d_{\text{emp}}$ is .80, which is smaller than the criterion value $\delta_{\text{crit}} = 1.0$. One may conclude that it is too small for the (psychological) prediction to have shown up. One may argue that there always is a (sometimes considerable) random variation and that the effect is rather large according to the conventions Cohen (1988) proposed (which, however, one should rarely use, if ever) and this may lead to the judgement that the prediction agrees with the data, despite the difference between empirical and critical effect size. – In addition to the random fluctuation, a systematic variation is usually present in the empirical effect sizes due to "variabilities of the particular measures and experimental manipulations", and thus: "Effect size estimates are helpful but must be interpreted with caution" (Cortina & Dunlap, 1997, p. 170).

Since psychological hypotheses are usually examined for a particular and completely specifiable empirical system, but not for populations, there is no need to estimate population effect sizes. The comparison addressed above may use the effect sizes computed for the sample. But if it seems preferable to refer to estimated population effect sizes, there seem to be no good arguments against their use. Some colleagues may find it appropriate to construct confidence intervals around the effect sizes, and there seems no good reason to caution against this use of confidence intervals around effect sizes; formulae for this are given by Cortina und Nouri (2000, pp. 8-9). For reasons given above no further test of significance should be performed on the empirical and on the critical effect sizes to avoid an infinite regress.

If the statistical prediction is decomposed there are several comparisons between the two kinds of effect sizes. If all empirical effects are at least as large as the critical values, a positive evaluation of the psychological prediction should result. If some empirical values are smaller than prespecified, one may nonetheless evaluate the psychological prediction positively. Or, one compares the average empirical effect size with the average criterion effect size and bases one’s evaluation concerning the psychological prediction on this comparison in a similar way to the one presented above. If all empirical values are smaller than the criterion values of the effect size, the researcher may find reasons for a positive decision on his or her psychological prediction admitting that the magni-
tudes of statistical effects are smaller than expected or hoped for. On the other hand, the researcher may come to a negative conclusion about the psychological prediction if a single effect size is smaller than prespecified.

Table 4 contains the possible patterns of results for $\beta < .5$, each consisting as a comparison of the critical value of test statistic with the empirical one a comparison of the empirical effect size with the criterion value fixed in advance. The table gives an answer to the question in which way to evaluate a psychological prediction, given results of a statistical test and a value of the effect size. Whether one uses the words of Table 2 (e.g., „PP conditionally shows up“) or chooses other descriptions such as „the PP shows up, but with a smaller effect than hoped for“ or the like, would make no difference as long as the pattern is described properly with respect to the statistical test and the effect size.

The single psychological predictions are: PP-1: „Mean recall under low imagery (2.5) is not as high as recall under high imagery (5.7);“ PP-2: „Mean recall under high imagery (5.7) is better than under low imagery (2.5);“ PP-3: „Mean recall under high imagery (5.7) is as good as under low imagery (2.5);“ PP 4: „Mean recall under high imagery (5.7) is not the same as under low imagery (2.5).“ PP-3 and PP 4 are tested twice, namely by an unsquared statistic ($H_{0(u)}$) and by a squared statistic ($H_{0(s)}$).

How is Table 4 to be read? Let us consider PP-1 which is tested by an unidirectional $H_0$. The statistical test remains insignificant, and there are two forms of an insignificant result. That is, in one version the empirical value of the test statistic (TS) is smaller than 0, which may lead to the judgement to call the PP having shown up. A second result shows statistical insignificance, thus again speaking for the PP. But the results points into the wrong direction, which means that the value of the test statistic is greater than 0, but smaller than the critical value: Overall this pattern of results may lead to the conclusion that the PP conditionally shows up. The next result is that $H_1$ is accepted, but the effect („in the wrong direction“) is smaller than prespecified. The smaller effect leads support to the conclusion that the PP-1 conditionally does not show up. The last pattern is described by statistical significance and moreover by an empirical effect that is greater than the critical one. Overall the PP-1 does not show up under this constellation of results.
**Type of psychological hypothesis (PH) and type of statistical hypothesis (H₀, H₁) derived (≈) from it**

<table>
<thead>
<tr>
<th>PH-1 ≈ PH-1 ≈</th>
<th>PP-1 ≈ PP-2 ≈ PP-2 ≈</th>
</tr>
</thead>
<tbody>
<tr>
<td>H₀: α: µ₁ - µ₂ ≤ 0; PH-2 ≈</td>
<td>PP-1 does not show up</td>
</tr>
<tr>
<td>PH-3 ≈ PP-3 ≈</td>
<td>PP-3 shows up</td>
</tr>
<tr>
<td>H₂: µ₁ - µ₂ ≠ 0</td>
<td>PP-4 shows up</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type of psychological hypothesis (PH)</th>
<th>Type of statistical hypothesis (H₀, H₁)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PH: psychological hypothesis; PP: psychological prediction. TSₚₑₜ denotes the empirical and TSₚₑₜ the critical value of the chosen test statistic with prechosen α. In case of two-sided tests, there is a lower critical value [TSₚₑₜ], and an upper critical value [TSₚₑₜ]. The symbol “&gt;&quot; means “either – or&quot;, i.e. only one of the two alternatives is possible. ESₚₑₜ: value of the effect size using planning the study. ESₚₑₜ: empirical value of the effect size. 1) If the empirical value of the test statistic is negative, the empirical value of the effect size is also negative, and takes on a lower value than the positive critical value. 2) In case of a uni-directional statistical H₀, agreeing with the psychological hypothesis (PH-1) the empirical effect size can be of any size in the direction which is opposite to the (disagreeing) alternative hypothesis (H₁). 3) A positive value of the test statistic implies a positive value of the empirical effect size, which, however, is always smaller than the critical value if H₀ is retained. Thus, comparisons between effect sizes in square brackets mean that the empirical value of the ES follows from accepting the H₀ and does not give any additional information. – The relationships shown hold for β &lt; .5. – Only parametric tests are considered.</td>
<td></td>
</tr>
</tbody>
</table>

**Table 4:** Possible pattern of results, consisting of decisions about statistical hypotheses and a comparison between the criterion and the empirical value of the effect size, and their meaning for the decision about the psychological prediction derived from the psychological hypothesis.
I argue for prespecified criteria, but I am aware of certain situations which might be used as counter-argument. It is argued that it may be unreasonable to retain an $H_0$ with prespecified $\alpha$ if the empirical value of the test statistic differs only very slightly from the critical value (see Harris, 1991, pp. 376-377), as is the case with the example I once encountered where $t_{temp} = -1.997$ and $t_{crit}(.05;60) = \pm 2.000$. Even if you decide to accept $H_1$ the effect size – if the test had been planned properly – is so small ($d_{emp} = -.51$) when compared to the criterion ($\delta_{crit} = \pm .80$ for $\beta = .10$ in the case considered) that the decision concerning the psychological prediction may or should be „negative“ despite the possible acceptance of the conforming $H_1$, since the expected or hoped for effect does not show up in the data, the difference being about 40%. A negative decision concerning the psychological prediction usually should lead to deciding that the psychological hypothesis is not corroborated; and the same decision would probably result if you had retained the $H_0$ above.

In addition, the researcher should think about his or her connecting statistical concepts to psychological variables. Have the criteria of adequacy and exhaustiveness been satisfied when formulating the predictions? Do the statistical concepts, e.g. means, give the information necessary with respect to the psychological hypothesis? In which way does her or his examination differ from the procedures and side conditions chosen by those researchers who established the first tests of the theory? Does her or his empirical system indeed fulfill the requirements necessary to apply the theory successfully? Is the scale type of the observable dependent variable appropriate for considering statements about (expected) means and with respect to the theory? Are other auxiliary hypotheses problematic which are explicitly or implicitly used? And so on. Critical reflections on these and other points may lead to modifications on the decisions and evaluations, particularly on the psychological prediction.

Overall, the judgment of the psychological prediction is again based on methodological rules, essentially in the same way as the decision on the statistical hypotheses, but incorporating more information. On the whole however, these rules leave far more freedom to the researcher than when deciding on the statistical hypotheses or when derive the prediction.

2.3. Evaluation of the psychological hypothesis

The final step is the evaluation of the psychological hypothesis based on the preceding decisions and evaluations. Since the variables usually addressed by the psychologi-
cal hypothesis are non-observable, whereas the variables of the psychological prediction are observable this step again contains inductive elements: The meaning or scope of non-observable variables is wider than that of observable variables. Thus, the positive decision on the statistical prediction and the according evaluation of the psychological hypothesis, but this evaluation should not rest solely on the prior decisions and evaluations. Whereas the decisions about the statistical hypotheses, the statistical prediction, and the psychological prediction rely on the data at hand, the evaluation on the psychological hypothesis must encompass an answer to the question: Why have the predictions shown up? Did they show up because the psychological hypothesis is “true,” or did they show up because some nuisance variables unaware of operated which led to effects one thinks due to the variation of the independent variables? Or: Why have the predictions not shown up? Did they not show up because the psychological hypothesis is “not true,” or did they not show up because some other nuisance variables operated which mask the effects attributable to the psychological hypothesis? These questions belong to the area of experimental validity (cf. Cook & Campbell, 1979; Hager, 1987, 1998; Hager & Westermann, 1983; Westermann, 1987) which is no part of statistical tests, so the questions have to be answered outside statistical considerations before the evaluation of the psychological hypothesis.

If the researcher thinks that his or her psychological hypothesis holds for the empirical system examined despite uncooperative data she or he may judge accordingly, but it may be hard to communicate the justification of this evaluation to other members of the scientific community. In the same vein, she or he may decide that the psychological hypothesis does not hold or is not applicable to the empirical system despite positive decisions on the preceding levels. How can this evaluation be justified?

In order to call the psychological hypothesis „corroborated for a particular empirical system“ the validity of the investigation must be considered to have been high enough – even if perfect validity can never be achieved. Although considerations concerning the experimental validity have to guide the construction of the empirical system, since various techniques of control can only be applied during designing the study, the actual validity may not be of the same degree as had been thought of before the study. Any manipulation which has not been performed as intended and any modifications concerning the various decision rules after the study, affect validity and lower at least one aspect of it.
One must consider the many auxiliary hypotheses usually necessary when conducting an experiment. The fewer important auxiliary hypothesis which can be identified and which seem not tenable in the light of how the experiment was actually performed, the more unambiguous the decisions. But even the tenability of the auxiliary hypothesis that randomization was effective can rarely be tested, and if it can, this test only refers to one or to some aspects of equally distributing „all other factors“ not mentioned by the psychological hypothesis to the experimental conditions. We only know for certain that the tenability of this auxiliary hypothesis can be assumed the safer the higher the number of elements who (or which) could be randomized.

The general message of the preceding considerations is that many considerations usually banned into the discussion part of empirical papers should be addressed when deciding on the psychological prediction and especially when deciding on the psychological hypothesis.

Overall, both evaluations of the psychological hypothesis (and many kinds of evaluations between the two „extremes“) seem defensible whatever the data says, and it is up to the researcher to convince the audience that her or his evaluation of the psychological prediction and of the psychological hypothesis is reasonable or justified – because or despite of the data. What is needed is „informed reasoning“ and „thoughtful decisions“ when evaluating the psychological assertions including the psychological predictions. Thoughtful decisions and evaluations cannot be enforced by empirical data, but empirical data may give the basis on which decisions and evaluations must be grounded in an empirical science.

Thus, we decide to call our (psychological) hypotheses corroborated if the empirical data give good support to this statement and if we control the risks of erroneous statements. We know that these evaluations refer to a certain experimental context and to our methods of scrutinizing and to our subjects and to our operationalizations of the non-observable constructs and so on. Changing the experimental context and/or our methods (using the U test instead of the t test, for example) may well lead to a different decision. Examining the imagery hypothesis with older adults may show that it does not hold for them.

Psychological hypotheses cannot be „corroborated“ or „not corroborated“ „overall“ („Overall, the hypothesis is valid“); instead, it is important that we identify as many situations, methods, subjects etc. as possible for which the hypotheses hold and for which they do not hold. The „generalization“ of the results should always pertain only
to systems which are „similar“ to the ones examined. What the researcher calls „similar“ cannot be discussed here (see Westermann & Gerjets, 1994; Westermann, Heise & Gerjets, 1992). This incorporates the notion, however, that we deal with a psychological hypothesis applicable to certain sets of empirical systems which share important similarities as if it were „true,“ i.e. in subsequent investigations we may act or behave accordingly and – e.g. – use the hypothesis as an auxiliary hypothesis necessary for examining another hypothesis.

3. Summary

In the present article the distinction between psychological and statistical hypotheses (Hager, 2000) is taken up and a four-step scheme is proposed for deriving predictions: Beginning with the psychological hypothesis refering to nonobservable constructs a psychological prediction refering to an experimental design and to the operationalizations of the theoretical terms in the psychological hypothesis is derived. In the next step, a statistical prediction refering only to statistical concepts is derived. The last step consists in deriving directly testable statistical hypotheses ($H_0$ and/or $H_1$) from the statistical prediction, that means that the last level only consists of at directly testable statististical hypotheses ($H_0$ and/or $H_1$). During the steps of derivation the two criteria of adequacy and exhaustiveness should be attended to, which are proposed to ensure that the possible empirical results are divided into two sets with opposite meanings, one set encompassing the result that agree with the psychological hypothesis and the other disagreeing with it. This division is central for all further judgements. The derivation is mainly an act which can be reconstrued by means of logic. – Two kinds of criteria are considered, one refering to the number of partial predictions which have to show up to make a positive decision on the statistical prediction and the other refering to the components of a decision: statistical tests for deciding on the statistical prediction and statistical tests and measures of effect size for evaluating the psychological prediction. A strategy is proposed which systematically incorporates effect sizes when evaluating the psychological prediction (and about psychological hypothesis). The conclusion part of this strategy is far less cogent or strict than the derivation part., since there is no logically valid form of inductive reasoning.
References


