Quality & Quantity (2008) 42:53–73
DOI 10.1007/s11135-006-9036-4


# All Bachelors are Unmarried Men ($p < 0.05$)

GEIR SMEDSLUND
*Norwegian Knowledge Centre for the Health Services, Postbox 7004 St. Olavs Plass, N-130 Oslo, Norway. E-mail: geir.smedslund@kunnskapssenteret.no*

**Abstract.** This paper adds to the list of criticisms against null hypothesis significance testing (NHST). I argue that when researchers do not analyze the conceptual relations among their variables, they may fail to distinguish between logical implications and empirical relations. It does not make sense to use significance testing on hypotheses involving conceptually related phenomena. The widespread lack of conceptual clarification also leads to very small effect sizes in psychology because it causes study participants to understand the stimulus material in different ways. Therefore, they answer in an inconsistent way. Researchers show an extremely low degree of ambition when they seek to show that psychological phenomena differ from chance, or when they try to disprove a hypothesis claiming that a psychological phenomenon does not exist. I see significance testing as a poor solution to the problem of tiny effect sizes in psychology. I recommend that psychological researchers be more explicit both about their main hypotheses and their auxiliary hypotheses. As examples, I analyse all quantitative articles in Issue 1, 2005 of the Journal of Health Psychology.

**Key words:** significance testing, nonempirical, conceptual analysis, pseudoempirical, quantitative analysis.

## 1. Introduction

My title alludes to Cohen's (1994) article "The earth is round ($p < 0.05$)", in which he raised some serious concerns about the practice of null hypothesis significance testing (NHST). Cohen and many others have criticised NHST for various reasons. I believe that I can add a new item to this list of criticisms, namely that *one should not perform significance tests on research questions involving conceptually related phenomena*. To my knowledge, no one has previously looked into the epistemological dimension of the practice of NHST.

### 1.1. THE DREAM OF LARGE EFFECT SIZES

Imagine a psychological science where predictions are almost perfect. Let's say that the researchers conducting a specific study can tolerate to be correct *only* 96% of the time although they hope for at least 99%. To contemporary mainstream psychological researchers this sounds wholly

utopian. They are used to very small effect sizes in their research. In fact they can usually tolerate, e.g., correlations as small as 0.05 as long as they are statistically significant. In other words, they are quite content with a situation where one of their variables predicts 2.5% of the variation in another one. Below I will show that the example with 96–99% predictive power is *not* utopian. It is quite possible for researchers who follow the suggestions given in the remainder of my paper.

### 1.2. THE RESEARCH PRACTICE OF MAINSTREAM PSYCHOLOGICAL RESEARCH

It seems that whenever mainstream psychological researchers want to test a hypothesis, they (1) automatically collect data; (2) automatically perform NHST to ascertain whether their results differ from pure chance (disproving that the relations under investigation are exactly zero); and (3) extensively cite other researchers who also have demonstrated results that differ from chance (Smyth, 2001).

### 1.3. MY ALTERNATIVE

I suggest that the widespread use of significance testing is a poor solution to the situation in psychological science with very small effect sizes. Further, I argue that the small effect sizes result from a lack of conceptual clarity. And this lack of conceptual clarity is a result of a failure to distinguish between conceptual relatedness and empirical regularities.

## 2. Conceptual Relatedness vs. Empirical Regularities

A presupposition for using significance tests is that we test *empirical* statements. An empirical statement is possibly true and possibly false. The only way to establish its veracity is to collect data. A necessary criterion of an (empirical) hypothesis is that its constituent terms are logically independent. A tautology (or near-tautology) cannot be a hypothesis because its terms are conceptually dependent. An example of such a tautology in psychology is: 'a surprised person has experienced something unexpected.' This cannot be a hypothesis because the meaning of the concept 'surprise' depends on the meaning of the concept 'expectedness.'

### 2.1. MAIN HYPOTHESES AND AUXILIARY HYPOTHESES

Research always involves more than just the main substantive theory of interest. As Meehl (1997) argued, one also has to take into account 'auxiliary theories relied on in the experiment...Ceteris paribus clauses ('other things being equal')... instrumental auxiliaries (devices relied on

for control and observation)...Realized particulars (conditions were as the experimenter reported)' (p. 398). When the results of a psychological study do not come out as the hypothesis predicted, the typical interpretation is that the hypothesis was wrong. An alternative interpretation is that one or more of the auxiliary conditions were not fulfilled.

### 2.2. RESPONSES ON QUESTIONNAIRES ARE NOT RANDOM

The reliability of a measuring instrument is improved if several *independent* ways of measuring a parameter give similar results. But in psychological inventories, the items are highly conceptually interdependent (Smedslund, 1987b). When psychologists conduct, e.g., questionnaire studies, they must silently take for granted that their subjects understand the instructions exactly as intended (including the meaning of the words making up the instructions). They also presuppose that they themselves understand the responses of their subjects correctly.

But the respondents understand that because they answered 'yes' to item **x**, they must answer 'no' to item **y** in order to be consistent. Thus, the researchers make up items that sound similar in order to measure a latent variable. In doing so, they have to assume that the respondents also share this perception of similarity. But the researchers also have to assume that the respondents do not perceive the items as too similar if the reliability coefficient shall have any purpose. There is no point in, e.g., testing the significance of a correlation between reporting of migraine and reporting of headache among a group of subjects. The null hypothesis is for instance: 'Knowledge that a subject reported migraine gives no knowledge as to whether the same subject reported headache.'

In other words, the responses of subjects on questionnaires are empirical regularities, but the *interpretation of* these responses must be that subjects must respond in a certain way as members of a society. There must be logical relations among the responses of the subjects. If not, the responses cannot be understood by others. It is not correct to test whether a logical relation is different from zero.

### 2.3. ARE WE TESTING THEORIES AND HYPOTHESES WHICH CANNOT BE REFUTED?

I believe that many psychological theories and hypotheses are tautologies or near-tautologies. Data can neither strengthen nor refute them. When data are in accordance with a theory, researchers tend to report that the theory was strengthened. But when data are not in accordance with the theory, the researchers do not abandon the theory but produce all kinds of ad hoc explanations for why the data did not agree with the theory (Ogden, 2003).

## 3. The Extra-Sensory Perception Paradigm

Researchers in the field of extra-sensory perception (ESP) try (among other things) to establish empirical evidence that human organisms are capable of communicating without using the five senses (telepathy). A common experimental design is the *ganzfeld procedure* (Bem, 1996). A sender sits in a separate soundproof room and concentrates on the 'target,' a randomly selected picture or videotaped sequence. Another person (the receiver) is deprived of any means for communicating with the sender using the senses. The receiver's task is to guess which target the sender is concentrating on. If both the sender and the receiver have sufficiently strong ESP, the receiver will in principle know every time which target the sender is thinking about. Since there is no documented evidence that this has ever happened, ESP researchers have a much less ambitious goal of showing that some dyads can 'beat' chance. If ESP is nonexistent, the receiver will make the correct guess on average in **1/n** of the trials if there are **n** different targets and if the number of trials is infinite. Likewise, a receiver with some degree of ESP will make the correct guess with a proportion of $\mathbf{p} > \mathbf{1/n}$ over an infinite number of trials. By chance, a sender/receiver dyad in the absence of ESP might communicate the correct guess with a proportion of $\mathbf{p} > \mathbf{1/n}$ over a finite number of trials. There is documented evidence that this has happened (Storm and Ertel, 2001). Because of this, there is no way of proving or disproving the existence of ESP. But one can utilise probability theory to estimate the probability of finding a certain amount of deviation from $\mathbf{p} = \mathbf{1/n}$ if the dyad has no ESP. The assumption is that the population proportion has a known distribution (e.g. binomial) around the proportion of **1/n** such that as the deviations become progressively larger and larger (on either side) these results become less and less probable. The probability is for example extremely small that the receiver will have no correct guesses over a large number of trials in the case of no ESP. On the other hand, the probability is virtually zero that a dyad with no ESP will perform perfectly.

ESP is, sympathetically viewed, a minuscule part of psychology, and, less sympathetically, not a science at all. But I take it as an example because I think that most research in psychology uses exactly the same paradigm, i.e., that psychological phenomena are hard to detect because of high levels of 'noise' in our measuring instruments and because of chance variations. As a consequence, what is really questioned is whether psychological phenomena and/or their inter-relations really exist!

## 4. The Presupposition that Psychology is an Empirical Science = One Must Always Collect Data

In this presupposition, researchers seem to be unaware of all the semantic constraints of their language. These constraints dictate what can and cannot be meaningfully stated. Any hypothesis or theory must be spelled out in the language of the researcher. Some hypotheses make sense intuitively, and some even are necessarily true. It would for example be difficult to find someone who would disagree with the statement: '*all sisters are female*.' On the other hand, almost everyone would disagree with the statement: '*all sisters are male*' or '*my sister is a boy'*. The reason why we intuitively agree/disagree with these utterances is that the construct 'sister' is so basic. In fact, one of the defining characteristics of 'sister' is 'female.' It would not be reasonable to collect data in order to find out whether all sisters are in fact female or to test whether the relation between sisterhood and female sex is exactly zero!

### 4.1. AS MEMBERS OF SOCIETY WE HAVE KNOWLEDGE ABOUT PSYCHOLOGICAL MATTERS

Fritz Heider, in his seminal work *The psychology of Interpersonal Relations* (Heider, 1958), was probably the first to pay attention to conceptual dependency in modern psychology. Later, Jan Smedslund has repeatedly argued that the main hypotheses of psychological theories are stated in such a way that they cannot be disproved (or strengthened) by data (Smedslund, 1978, 1987a,b, 1991, 1997b, 1999, 2002). Some of the psychological theories and models which have been shown to be nonempirical include: Banduras theory of self-efficacy, Thorndike's law of effect, Seligman's theory of learned helplessness, Bateson's double-bind hypothesis, Festinger's dissonance theory, Kelly's theory of personal constructs, the Health Belief Model, Prochaska and DiClemente's transtheoretical theory of change (stages of change) and the Theory of Planned Behaviour. A number of writers have partly or fully agreed with these thoughts (Ogden, 2003; Kukla, 1989; Ossorio, 1991; Parrot and Harré, 1991; Shotter, 1991; Shweder, 1991; Wallach and Wallach, 1998a,b, 1999), but still, most mainstream psychologists seem to implicitly presuppose that no knowledge can be gained without collecting data. I believe that the reason is that most researchers have not analysed their constructs in such a way that they can even know whether the constructs are conceptually independent or not.

### 4.2. PSYCHOLOGIC

Jan Smedslund has constructed (Smedslund, 1988) and refined (Smedslund, 1997b) a system called Psychologic, a kind of psychological calculus

comprised of a small number of primitive (undefined) terms, definitions of technical terms, axioms and a larger number of derived theorems and corollaries. Psychologic is not a formal logical system but a system of conceptual relations as they are embedded in everyday language. Whereas proofs in formal logic are independent of concept meaning, conceptual logic relies on the meanings of concepts.

Three philosophical distinctions are important in this regard. First, *analytic* propositions are derivable from the meanings of the constituent terms, while *synthetic* propositions are not. Second, the truth-value of *empirical* propositions is only knowable through experience, while knowledge that is *a priori* is knowable *other than* through experience. Third, *contingent* propositions are possibly true and possibly false, while *noncontingent* propositions are necessarily true or necessarily false. The proposition: '*all sisters are female*' is analytic and a priori while the proposition: '*fighting spirit is positively related to the number of natural killer cells*' is synthetic and empirical. I should add that not all propositions fall clearly into one or the other category (Putnam, 1975). And it is always the case that a given proposition is only empirical or a priori in relation to a set of premises. '*My sister is a boy*' is analytic (and false) in isolation, but if we assume that my sister has undergone a sex change operation, the proposition becomes empirical. My point here is that psychological researchers often are not explicit enough in order to clarify the epistemological status of their hypotheses.

As for the modal distinction between contingent and noncontingent propositions, I believe that psychological common sense is a priori and contingent. This pertains to what we know about being human because we are humans. This knowledge is a priori, but it is contingent because it could have been different if humans were different. My example '*All sisters are female*' is true in all possible worlds, that is, noncontingent. However, '*a person will tend to remember what he or she takes to be possibly relevant for the achievement of his or her goals*' is a priori but contingent, because it could be different in a world where persons are different. This a priori knowledge is not tautological. It tells us something about the world.

The propositions of Psychologic are claimed to be consensually self-evident, meaning that competent speakers of a language will agree that they are true and that their negations are senseless or contradictory. This consensus has in fact been empirically tested in speakers of eight different languages (Arabic, English, Ewe, Norwegian, Tamil, Turkish and Vietnamese) (Smedslund, 1997a) and Urdu (Smedslund, 2002). The average agreement about the status of the propositions in Psychologic as being always true was about 97%, and when the exceptions were studied more closely, many could be seen to result from misunderstandings. Researchers are concerned about the predictive power of their theories. The predictive power of Psychologic is far better than any psychological theory

that I know of. I stated in the introduction that a predictive power of 96% is not utopian, and Psychologic is an example of this high degree of predictive power.

### 4.3. PSEUDO-EMPIRICISM

Scientific studies always involve main hypotheses and auxiliary hypotheses, and a result in accordance with an empirical main hypothesis means that this hypothesis *and* all the auxiliary hypotheses are supported (Smedslund, 1995). If the result is not in accordance with the main hypothesis, this hypothesis and/or at least one auxiliary hypothesis must be wrong (Meehl, 1997). On the other hand, a result in accordance with, or contrary to, a nonempirical main hypothesis has no bearing on its truth-value, but only strengthens the auxiliary hypotheses or weakens at least one of them.

Studies that collect data in order to investigate the truth-value of nonempirical hypotheses have been labelled *pseudoempirical* (Smedslund, 1984).

### 4.4. HYBRID STUDIES

In a wider sense, health psychology also involves hybrid studies in which one is interested in relations between subjective and biological variables. One specific example is the hypothesis about a relation between helplessness and cancer. These variables do not seem to be conceptually related, so the only way to study this hypothesis is to collect data. But researchers in psychoneuroimmunology also use significance tests, and they are not immune to pseudo-empiricism! It would for example be incorrect to study how perceived control moderates the relation between helplessness and natural killer cell counts because the relation between perceived control and helplessness can be shown to be nonempirical. In the Internet-based encyclopedia The Free Dictionary (2005), one of the meanings of the word helplessness is: 'a feeling of being unable to manage.' A person being in a state of having maximal control over something $X$, and at the same time having minimal ability to manage $X$ does not make sense. As persons, we know that having control is the same thing as being able to manage. Conceptual analysis is the appropriate method for studying relations among subjective phenomena.

## 5. What Significance Testing is and is not

Without having any prior knowledge, the sample mean of a parameter is the best estimate of the population mean (especially when assuming that the parameter is normally distributed in the population). Two randomly drawn sample means will differ only by sampling variation with known

characteristics. A significance test provides the posterior likelihood for the data conditional on the truth of the hypothesis tested (usually the null hypothesis). As a consequence, significance testing is *not* appropriate when samples are either not randomly drawn from the population or when the parameter is not normally distributed in the population. Furthermore, the test does not give direct information about the conditional probability of the hypothesis given the data (the posterior likelihood of the hypothesis). And finally, the researcher always has some idea about the probabilities of the null hypothesis (its prior probability). Usually the researchers do not believe in the null hypothesis. Despite of this, they are obliged to retain the null if the test does not come out significant.

Significance tests should only be used in situations in which there is incomplete information about an empirical relation such that every conclusion must involve probability statements. We have, e.g. two samples with different means on a variable. We want to know whether this difference came about by chance or whether it was produced through some systematic process. If two samples were randomly drawn from a population in which the parameter of interest is normally distributed, the laws of probability would allow us to quantify the conditional probability of the difference in the sample parameter values being this large or larger. If the difference is very unlikely, we can choose between two rather different conclusions. We may simply conclude that a very unlikely occurrence has occurred, or we may conclude that the difference has other causes in addition to random sampling variation.

### 5.1. WHAT EXACTLY IS A NULL HYPOTHESIS?

Typically, the null hypothesis has been what Cohen (1994) labelled the 'nil hypothesis.' This is a hypothesis claiming that, e.g. a correlation or a difference between two means is literally zero (nil). Since a point value hypothesis can never be literally true, 'zero' has commonly been understood as a region around zero so small that it is of no practical value.

But there are other forms of null hypotheses. In other sciences, such as Physics, a theory sometimes can make point value predictions as opposed to psychological theories, which usually only make directional predictions. Seen in this way, a null hypothesis can be, e.g. that the theory predicts a mean score on a variable to be 20. The significance test seeks to advice us whether the observed (sample) value is highly improbable given that the true (population) value is 20.

Significance testing has repeatedly been severely criticised since its inception in the 1930s (Berkson, 1938; Jones, 1955; Kish, 1959; Rozeboom, 1960; Meehl, 1978; Oakes, 1986; Cohen, 1994; Schmidt, 1996; Hubbard et al., 1997; Sterne and Davey Smith, 2001), and I will not repeat these

criticisms in detail here. But to my best knowledge, none of the authors treated the point that I raise in this paper, namely that significance testing is inappropriate when there is conceptual dependency among the studied variables.

### 5.2. MAINSTREAM PSYCHOLOGISTS' VIEW OF THE SCIENCE

Psychology appears extremely complicated. No relations seem to be universally true. In nearly every given area, there is significant and non-significant conflicting evidence. Effects are usually small. Even when testing a hypothesis which seems intuitively highly plausible, the results are often negative. Seen from this background, the existence of generally valid and necessarily true psychological hypotheses appears highly unlikely. There is so much noise in the data that it is necessary to have a way of separating whatever systematic processes might be there from all the noise. If a researcher believes that all psychological hypotheses are empirical, it follows that nothing is known about a relation before data are collected. For instance, perceived control over a threat may make you more or less anxious. Or perceived control may be unrelated to anxiety. We need to collect data to find out. If a relation is absolutely unknown, it may seem like a good start to determine whether it is significantly different from zero or not, and its direction.

### 5.3. AN ALTERNATIVE VIEW

By 'psychology' I mean the science of subjective processes (what exists FOR persons, e.g. reflective and unreflective awareness of wants, beliefs, feelings and acts). If psychology is looked upon with the empirical/a priori distinction in mind, the field appears as a simple and orderly network of concepts. Relations follow necessarily from other relations.

Strong effect sizes rely on high reliability of study variables. In questionnaire studies this means that (1) each respondent has to answer the study questions in a consistent way, and (2) different respondents have to answer in ways that make them systematically different. But in order to do this, they must understand in the same way what the questions mean and the implications of the different answers. I will deal with each point separately:

(1) If the respondents do not understand the intended meanings of the questions, they will not understand the implications of the questions. Because of this, they will not respond in a consistent way from question to question. Example: "If I answer 'yes' on answer *X*, I have to answer 'no' on question *Y* and 'yes' on question *Z* in order to be consistent." Any within-person statistic will be downwardly biased in this case.

(2) If the respondents do not understand the questions in the same way, they will answer in a more diverse way. John thinks that question *A* means *X* and Joan thinks that question *A* means *y*. In this case any between-person statistic will get artificially reduced.

In order to make my point clear, I give some examples. I chose the (at the time) most recent issue of a leading journal in the field in which I have the most content expertise (Journal of Health Psychology). From this issue (Marks, 2005); I included all articles with a quantitative analysis. There were six such articles (Bagozzi et al., 2005; Ingledew et al., 2005; Insel et al., 2005; Iwasaki et al., 2005; Mccabe and Judicibus, 2005; Rivers et al., 2005). Table I gives a summary of the studies. Because of space limitations, I could only take a few examples from each article to show that my worries about current research practice are upheld after close reading of a journal issue published in 2005.

## 6. The Quantitative Articles of Issue 1, 2005 in Journal of Health Psychology

### 6.1. BAGOZZI ET AL.

This article studies the actions and interactions of *Pharmacy and Therapeutic Committee* groups. These committees are multidisciplinary teams of physicians, pharmacists, nurses and other health professionals that evaluate and select medications for inclusion in the hospital's formulary. The researchers used the *Social Relations Model* (Kenny, 1994) to partition the variance in group members' behaviour into five components: (1) actor effect; (2) partner effect; (3) relationship effect; (4) interaction effect; and (5) residual error.

The researchers report a large number of results about interpersonal behaviour in the group which cannot be detailed here, but one was that the mean level of cooperation given from the physician chairs to physician members was 4.09 on a 5-point scale. The corresponding level of cooperation given from physician members to administrators was 3.46 (their Table 5, p. 53). These are empirical results. In their Table 6 (p. 55), they tested whether the effects were significant. I take this to mean that they tested whether the effects are different from zero. Stated bluntly, they tried to disprove whether persons who sit together in a group for hours have no influence on each other's behaviour whatsoever (the null hypothesis).

They also have some other examples of findings which seem to be non-empirical. For example they reported that: ' ... chairs who tend to be frustrated and angry in their reactions towards others also elicit frustration and anger from these others' (p. 55), and ' ... the more one particular actor

*Table I.* Summary of the included quantitative articles from Issue 1, 2005 of the Journal of Health Psychology

| First author | Theories and models | Concepts | Hypotheses/ main points | Design and Methods |
|---|---|---|---|---|
| Bagozzi | Social Relations Model, Interdependence theory | Cooperation, influence, frustration, enjoyment | How do the members of Pharmacy and Therapeutics Committee members influence each other? Does elicitation and reception of cooperation, influence, frustration and enjoyment vary by role? | Survey with round-robin design |
| Ingledew | Value-expectancy model of goal commitment | Perception, work-related goals, affective well-being, goal commitment, success expectation | (1) Goal value and goal success expectation would both positively influence goal commitment; (2) value would positively influence positive affects; (3) Success expectation would negatively influence negative affects; (4) Commitment and affects would not directly influence each other; (5) Value would be influenced by competition, conflict (negatively), personal origin, publicness and specificity; (6) Success expectation would be influenced by ability, complexity (negatively), control, difficulty (negatively), feedback, support, time and tools; (7) these determinants of value and success | Structural equation modeling |

*Table I.* continued

| First author | Theories and Models | Concepts | Hypotheses/ main points | Design and Methods |
|---|---|---|---|---|
| | | | expectation would not have direct effects on commitment, positive affects or negative affects | |
| Insel | Common sense model of self-regulation | Illness representation, breathing, breathlessness | Determine if there are important differences in the underlying organization (illness representation) of common concepts associated with breathing and breathlessness between common sense experts (patients with COPD or asthma) and expert health care providers, pulmonologists and advance practice nurses (pulmonary nurse specialists, PNS) | Visual analogue scales |
| Iwasaki | | Leisure participation, coping, relaxing leisure, stress, social leisure, cultural leisure | Type of leisure activity matters in predicting immediate adaptational outcomes (coping effectiveness, coping satisfaction and stress reduction) and mental and physical health | Repeated measures |
| McCabe | | Psychological well-being, quality of life, economic disadvantage | Evaluate the impact of economic disadvantage among people with multiple sclerosis (MS) on their psychological well-being and quality of life | questionnaire |

*Table I.* continued

| First author | Theories and models | Concepts | Hypotheses/ main points | Design and Methods |
|---|---|---|---|---|
| Rivers | Prospect theory | Intentions, behaviour, negative affect | Loss- and gain-framed messages differentially influence health behaviours depending on the risk involved in performing the behaviour | Randomized experiment |

showed enjoyment and satisfaction toward the specific partner, the more the partner reciprocated enjoyment and satisfaction' (p. 56).

A problem with significance tests is that trivial results tend to become highly significant if sample size is large enough. In their Table 8 (p. 59), variance estimates (for dyadic reciprocity outcomes) as small as 0.01 are reported as significant. Another problem is that a sharp alpha cut-off level (usually 0.05 or 0.01) makes one ignore that results fall on a continuum. In their Table 8, it seems that variances of 0.01 are regarded as important but 0.00 (with two significant digits) are not.

I think that the Social Relations Model is a model that cannot be tested because it is nonempirical. Therefore, using structural equation modelling (SEM) and estimating model fit is pseudoempirical. Put in another way, significance testing of the model fit provides the conditional probability of the data given that the model is true. But we know beforehand that the model is true. An interesting point is that logic (or common sense) is ordinal in the sense that it, at its highest level of precision, only deals with relations such as 'larger than' or 'less than.' You can say things like '*the stronger you believe that you can achieve a desired goal, the harder you will try*.' But you cannot meaningfully state that if your belief increases 2.4 times, you will try 2.4 times as hard. But empirical (and pseudoempirical!) research makes predictions on an interval or ratio scale.

These authors seem to have presupposed that the Social Relations Model is a theory that can be refuted if the data do not support it, i.e. they believed that it is an empirical model. Moreover, they aimed to show that the influence of one person on other persons in a group is larger than zero. In doing this, they proceeded exactly like ESP researchers in trying to show that the observed relations differ from chance.

### 6.2. INGLEDEW ET AL.

The aim of this paper was to clarify how perceptions of work-related goals influence affective well-being and goal commitment.

Ingledew et al., revealed that they were unaware of the empirical–nonempirical continuum in the very first sentence of the article: 'The subjective work environment is an important influence on well-being' (p.102). This can be shown to be nonempirical because 'subjective' has to refer to how you perceive, experience and feel, and well-being also is about how you feel. With some substitution, the sentence can be re-written as: 'How you feel about the work environment is an important influence on how you feel' (a truism). But Ingledew et al., included a citation here as if the proposition needed empirical justification.

The authors tested a model using structural equation modelling. The model tested whether 'individuals are committed to a goal as a result of valuing it or expecting to succeed at it or both' (p. 102). The authors perceivably took this as an empirical statement. But this implies that (1) the likelihood that a person will *try to achieve* a goal is (conceptually) unrelated to whether the person *expects to succeed* at achieving the goal and (2) the likelihood that a person will *try to achieve a goal* is (conceptually) unrelated to whether the person *values the goal*. But no one would argue that (1) or (2) make sense as stand-alone propositions. Using significance testing in this case means trying to disprove that there is no relation between trying to attain a goal and believing that it is possible to attain the goal. Another example: complexity and difficulty of the goal correlated positively ($r = 0.66$, $p < 0.01$). In this case the null hypothesis assumed that there is no relation between how complex a goal is and how difficult it is to achieve!

### 6.3. INSEL ET AL.

The researchers wanted to determine

> "if there are important differences in the underlying organisation (illness representation) of common concepts associated with breathing and breathlessness between common sense experts (patients with COPD or asthma) and expert health care providers, pulmonologists and advance practice nurses (pulmonary nurse specialists, PNS)" (p. 151).

Two patient groups, those with chronic obstructive pulmonary disease (COPD) and those with asthma, and two provider groups, pulmonologists and nurse specialists (PNS), rated the dissimilarity between each of 55 pairs of concepts on visual analogue scales (VAS).

It seems that the different ways patient or provider groups conceptualise a disease is largely empirical. Hence, I have no problems with collecting data to examine this. However, there seems to be some nonempirical elements in their data. On page 157 they wrote: "*Breathing effort* and *work of breathing* are closely related for participants in all four groups, suggesting that these concepts are not distinguishable and not useful … " In this case, it seems that the authors drew this conclusion based on the *data* that they had collected. But my alternative interpretation is that the respondents saw the two phrases as more or less synonyms. In other words, the relation between 'work of breathing' and 'breathing effort' is not empirical.

Their Table 5 (p. 157) shows similarities in how the groups linked (perceived as similar) concepts. They performed a significance test which returned a probability of less than 1% of shared links based on chance alone. Their null hypothesis was apparently that persons in different occupations have absolutely no common conceptualisation of the world!

In the discussion they wrote (p. 159): "These findings support the more general hypothesis that the dissimilarity judgments made by each group of participants reflect their subjective experience with pulmonary disease and their training." In short, they claimed that their data showed that people's judgments are influenced by their subjective experiences. But it makes no sense to deny this. No one would agree with the assertion that (other things equal), people's judgments are wholly unrelated to their prior experiences.

### 6.4. IWASAKI ET AL.

> "The purpose of this study was to examine whether and the extent to which leisure predicts effective coping with stress and good physical and mental health over and above the effects of general coping – coping that is not directly associated with leisure" (p. 83).

They stated the following research questions: (1) Does the type of leisure activity matter in predicting better adaptational outcomes? (2) Do the enjoyment indicators of leisure significantly predict positive adaptational outcomes over and above the effects of the frequency indicators of leisure?

The researchers indicated that they were not aware of the distinction between empirical and nonempirical when they stated that: "Evidence has been found that coping influences the relationship between stress and illness or health" (p. 80). The relation between coping and stress is nonempirical. The most commonly accepted definition of stress (mainly attributed to Richard S. Lazarus) is that stress is a condition or feeling experienced when a person perceives that demands exceed the personal and social resources the individual is able to mobilise. It does not make much sense to claim that "*the demands exceed my resources but I am coping fine*."

Their Table I (p. 88) shows 276 correlations. Probability theory tells us that when 20 correlations are computed on variables which are unrelated in the population, one of them will be significant at the 5% alpha-level by pure chance in a randomly drawn sample from this population. Therefore, it is common to adjust for this. The Bonferroni correction in this instance would be 0.05/276 = 0.00018, which is the alpha-level distinguishing significance from nonsignificance. The authors did not report any such adjustment. This means that the number of significant correlations expected by chance is 14. They reported 113 significant correlations, which is way above the expected number. This is clear evidence that the population correlations are not zero. Take for instance the correlation between stress level and mental health at time 2. This correlation is $-0.59$ which they report as significant at $p < 0.05$. This means that the more stress a person experiences, the worse is the person's mental health. But the researchers implicitly acted as if they believed that there might possibly be no relation between a person's stress level and the person's mental health. They estimated the probability of finding a correlation of, or larger than $-0.59$ if the real correlation between stress level and mental health were 0.00. The authors did not define 'mental health' in such detail that they were able to deduce what good mental health would logically imply about the level of stress.

### 6.5. MCCABE ET AL.

This study investigated the impact of economic disadvantage among people with multiple sclerosis (MS) on their psychological well-being and quality of life. Psychological well-being was described as involving (lack of) depression, anxiety, confusion and fatigue. Quality of life had four subdomains: physical health, psychological, social and environmental domains. Information was obtained on income, lost income, costs of MS, economic pressure, coping, psychological well-being and quality of life. Economic pressure involved whether respondents felt they could not make ends meet and whether they had money left over at the end of the month. Coping was subdivided into problem-focused coping, detachment, wishful thinking, seeking social support and focusing on the positive.

In the discussion (p. 172), the authors indirectly seemed to acknowledge that some relations are nonempirical: "*Not surprisingly*, high levels of social support coping were also associated with better functioning on the social domain, and high levels of detachment were associated with lower levels of engagement in activities within the person's broader environment" (Italics added). Is it possible to function worse socially if you receive informational support, tangible support and emotional support? And is it possible to be more engaged and more detached at the same time?

### 6.6. RIVERS ET AL.

This article used a randomised experiment to test the hypothesis that loss- and gain-framed messages differentially influence behaviours depending on the risk involved in performing the behaviour. They predicted (and found) that loss-framed messages emphasising the costs of not detecting cervical cancer early (a risky behaviour) and gain-framed messages emphasising the benefits of preventing cervical cancer (a less risky behaviour) were most persuasive in motivating women to obtain a Pap-test. Prospect theory (Kahneman and Tversky, 1990) suggests that people are willing to tolerate risks when thinking about the potential losses or negative consequences of a choice, but avoid risks when thinking about the potential gains or benefits of a choice. They did not seem to realise that the relations under study could be nonempirical.

In some cases they found only a 'borderline significance' of $p = 0.078$. Of course, this is really a nonsignificant result, and according to common practice they should have rejected the hypothesis. But they did not. Instead they seemed to uphold their belief in the hypothesis and explained the 'marginal statistical significance' in the following manner: "First, the effect of the interaction may have been less strong because the message had limited impact" (p.76). This seems to indicate that they thought that the reason the $p$-value did not fall below the 0.05-level was because some auxiliary hypotheses were not met (e.g. that the message had sufficient impact). In other words, they should have rejected the theory based on the results, but the theory can probably not be rejected because it is nonempirical and true. And the authors seemed to have unreflectively acknowledged this.

## 7. Concluding Remarks

In this paper, my aim is to add to the list of criticisms against NHST. I claim that many null hypotheses in contemporary psychological research are negations of propositions which can be shown to be nonempirical. I believe that this is the case because researchers are not explicit enough about defining their key concepts and their conceptual inter-relations. I also claim that lack of conceptual clarity leads to very small effect sizes in psychological research. These small effect sizes have lead to a lot of frustration. As a last resort, the research community has clinged to NHST as a means of showing that there *is* an effect even though it is tiny. Maybe it also is some consolation in labelling this tiny effect 'significant,' which means 'important,' 'notable' or 'influential.'

My simple recommendations to researchers are that hypotheses should always be subjected to a conceptual analysis to find out whether they are empirical or not. This requires increasing clarity. Can anyone argue against

clarity? If they are not empirical, a data collection can only test auxiliary hypotheses about, for example, reliability and validity. If they are empirical, some sort of effect magnitude and associated degree of precision should always be reported (e.g. correlation, standardised mean difference, odds ratio, number needed to treat, relative risk, absolute risk reduction).

In my examples from the Journal of Health Psychology, I could only highlight selected examples from each article. It could be argued that I did not give the articles full justice in this way. My response is that I can only see one way that the authors can respond to my critique. This is by defining every key concept and showing that they *are* conceptually independent. If they can show this, they should explain why significance testing is appropriate in each specific case.

In 2000, I had an open peer commentary article in the Journal of Health Psychology (Smedslund, 2000). In the target article, I argued that a number of theories and models in health psychology (Health Belief Model, Theory of Planned Behaviour and Social Cognitive Theory) are not empirical. Of 24 invited leading authorities and theorists, only four actually commented on my article. The rest chose to remain silent. The editor, David F. Marks, wondered why and offered two probable explanations (Marks, 2000b). The first was that theorists in health psychology are not used to philosophical and logical arguments. The second was that they fear that acknowledging my points would be a threat to the very foundation of (health) psychology. Issue 1, 2005 of the JHP was published almost 5 years after my open peer commentary article appeared in the same journal. Those 5 years seem to have passed without any detectable change in research practice among contributors to the journal.

In this brief article, many methodological, theoretical and philosophical problems could not be discussed in detail. In order to make my argument clear, I very briefly described the system of Psychologic and the significance test controversy. For more detailed discussions about Psychologic, the reader is referred to target articles with peer commentaries in Psychological Inquiry (Pervin, 1991), Scandinavian Journal of Psychology (Helstrup et al., 1999), and the afore-mentioned issue of Journal of Health Psychology (Marks, 2000a). Regarding the significance testing controversy, I recommend the book 'What if there were no significance tests' (Harlow et al., 1997) and the following papers (Cohen 1990, 1994; Falk, & Greenbaum 1995; Schmidt 1996; Rothman and Greenland 1998; Krueger 2001; Sterne et al., 2001).

The claims made in this paper may seem extreme to many readers. I have tried to show that these problems are abundant in a recent issue of an international psychology journal (with an impact factor of 0.8 in the Social Sciences Citation Index in 2002). In every article, I easily found examples of misuse of significance testing and pseudoempirical analysis. This could largely have been avoided if the authors had defined their key concepts and

analysed the epistemological and modal status of their main hypotheses. I have summarised the articles in Table I. In my view, most of the key concepts are not sufficiently defined, and the hypotheses have not undergone a conceptual analysis in any of the articles. Without a high level of conceptual clarity it is hard to distinguish between empirical and nonempirical research questions. It is, of course, quite possible that I have misunderstood some of the material in the articles. But if the authors of my example articles are to prove me wrong and show that their analyses really *are* empirical, they are forced to heighten their level of conceptual clarity. And that would make me happy!

## References


Bagozzi, R. P., Ascione, F. J. & Mannebach, M. A. (2005). Inter-role relationships in hospital-based pharmacy and therapeutics committee decision making. *Journal of Health Psychology* 10: 45–64.

Bem, D. J. (1996). Ganzfeld phenomena. In: G. Stein (ed.), *Encyclopedia of the Paranormal*. Buffalo, NY: Prometheus Books, pp. 291–296.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association* 33: 526–542.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist* 45: 1304–1312.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist* 49: 997–1003.

Falk, R. & Greenbaum, C. W. (1995). Significance tests die hard: the amazing persistence of a probabilistic misconception. *Theory and Psychology* 5: 75–98.

Harlow, L. L., Mulaik, S. A. & Steiger, J. H. (1997). *What if there were no Significance Tests?* Mahwah, NJ: Lawrence Erlbaum Associates.

Heider, F. (1958). *The Psychology of Interpersonal Relations*. New York: Wiley.

Helstrup, T., Rognes, W. & Vollmer, F. (1999). Psychologic and memory. *Scandinavian Journal of Psychology* 40(Suppl. 1): 1–138.

Hubbard, R., Parsa, R. A. & Luthy, M. R. (1997). The spread of statistical significance testing in psychology. The case of the Journal of Applied Psychology, 1917–1994. *Theory and Psychology* 7: 545–554.

Ingledew, D. K., Wray, J. L., Markland, D. & Hardy, L. (2005). Work-related goal perceptions and affective well-being. *Journal of Health Psychology* 10: 101–122.

Insel, K. C., Meek, P. M. & Leventhal, H. (2005). Differences in illness representation among pulmonary patients and their providers. *Journal of Health Psychology* 10: 147–162.

Iwasaki, Y., Mannell, R., Smale, B. J. & Butcher, J. (2005). Contributions of leisure participation in predicting stress coping and health among police and emergency response service workers. *Journal of Health Psychology* 10: 79–99.

Jones, L. (1955). Statistics and research design. *Annual Review of Psychology* 6: 405–430.

Kahneman, D. & Tversky, A. (1990). Prospect theory: an analysis of decision under risk. In: P. K. Moser (ed.), *Rationality in Action: Contemporary Approaches* vol. New York, NY: Cambridge University Press, pp. 140–170.

Kenny, D. (1994). *Interpersonal Perception: A Social Relations Analysis*. New York: Guilford Press.

Kish, L. (1959). Some statistical problems in research design. *American Sociological Review* 24: 328–338.

Krueger, J. (2001). Null hypothesis significance testing: on the survival of a flawed method. *American Psychologist* 56: 16–26.

Kukla, A. (1989). Nonempirical issues in psychology. *American Psychologist* 44: 785–794.

Marks, D. F. (2000a). A pragmatic basis for judging models and theories in health psychology: the axiomatic method. Open peer commentary. *Journal of Health Psychology* 10(1): 5–176.

Marks, D. F. (2000b). Editorial. *Journal of Health Psychology* 5: 131–132.

Marks, D. F. (ed.) (2005). *Journal of Health Psychology* 10(1).

Mccabe, M. P. & Judicibus, M. D. (2005). The effects of economic disadvantage on psychological well-being and quality of life among people with multiple sclerosis. *Journal of Health Psychology* 10: 163–173.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology* 46: 806–834.

Meehl, P. E. (1997). The problem is epistemology, not statistics: replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In: L. Harlow, S. Mulaik & J. Steiger (eds.), *What if There were no Significance Tests?* Mahwah, NJ: Erlbaum, pp. 393–425.

Oakes, M. (1986). *Statistical Inference: A Commentary for the Social and Behavioral Sciences*. New York: Wiley.

Ogden, J. (2003). Some problems with social cognition models: a pragmatic and conceptual analysis. *Health Psychology* 22: 431–434.

Ossorio, P. G. (1991). Naive baseball theory. *Psychological Inquiry* 2: 352–355.

Parrot, W. G. & Harré, R. (1991). Smedslundian suburbs in the city of language: the case of embarrassment. *Psychological Inquiry, 2:* 358–360.

Pervin, L. (1991). The pseudoempirical in psychology and the case for Psychologic. *Psychological Inquiry* 2(4): 325–382.

Putnam, H. (1975). The analytic and the synthetic. In: H. Putnam (ed.), *Philosophical Papers: Mind, Language and Reality*, vol. 2. Cambridge: Cambridge University Press, pp. 33–69.

Rivers, S. E., Salovey, P., Pizarro, D. A., Pizarro, J. & Schneider, T. R. (2005). Message framing and pap test utilization among women attending a community health clinic. *Journal of Health Psychology 10:* 65–77.

Rothman, K. J. & Greenland, S. (1998). Approaches to statistical analysis. In: K. J. Rothman & S. Greenland (eds.), *Modern Epidemiology*, 2nd edn. London: Lippincott Williams & Wilkins, pp. 183–199.

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin* 67: 416–428.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: implications for training of researchers. *Psychological Methods* 1: 115–129.

Shotter, J. (1991). Measuring blindly and speculating loosely: But is a 'Psychologic' the answer? *Psychological Inquiry* 2: 363–365.

Shweder, R. A. (1991). On pseudoempiricism, pseudodeductionism, and common sense. *Psychological Inquiry* 2: 366–370.

Smedslund, G. (2000). A pragmatic basis for judging models and theories in health psychology: The axiomatic method (target paper). *Journal of Health Psychology* 5: 133–149.

Smedslund, J. (1978). Bandura's theory of self-efficacy: a set of common sense theorems. *Scandinavian Journal of Psychology* 19: 1–14.

Smedslund, J. (1984). What is necessarily true in psychology? In: J. R. Royce and L. P. Mos (eds.), *Annals of Theoretical Psychology*. New York, London: Plenum Press, pp. 241–272.

Smedslund, J. (1987a). Ebbinghaus, the illusionist: how psychology came to look like an experimental science. *Passauer Schriften zur Psychologiegeschichte* 5: 225–239.

Smedslund, J. (1987b). The epistemic status of inter-item correlations in Eysenck's Personality Questionnaire: the a priori versus the empirical in psychological data. *Scandinavian Journal of Psychology* 28: 42–55.

Smedslund, J. (1988). *Psycho-Logic*. Berlin, Heidelberg, New York: Springer Verlag.

Smedslund, J. (1991). The pseudoempirical in psychology and the case for Psychologic. *Psychological Inquiry* 2: 325–338.

Smedslund, J. (1995). Auxiliary versus theoretical hypotheses and ordinary versus scientific language. *Human Development* 38: 174–178.

Smedslund, J. (1997a). Is the 'psychologic' of trust universal? In: S. Niemeier & R. Dirven (eds.), *The Language of Emotions. Conceptualization, Expression, and Theoretical Foundation*. Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 3–13.

Smedslund, J. (1997b). *The Structure of Psychological Common Sense*. Mahwah, New Jersey, London: Lawrence Erlbaum Associates, Publishers.

Smedslund, J. (1999). Psychologic and the study of memory. *Scandinavian Journal of Psychology* 40: 3–17.

Smedslund, J. (2002). From hypothesis-testing psychology to procedure-testing psychologic. *Review of General Psychology* 6: 51–72.

Smyth, M. M. (2001). Fact making in psychology: the voice of the introductory textbook. *Theory and Psychology* 11: 609–636.

Sterne, J. A. & Davey Smith, G. (2001). Sifting the evidence – what's wrong with significance tests? *British Medical Journal* 322(January): 226–231.

Storm, L. & Ertel, S. (2001). Does psi exist? Comments on Milton and Wiseman's (1999) meta-analysis of ganzfeld research. *Psychological Bulletin* 127: 424–433.

The Free Dictionary (2005). Retrieved December 21, 2005, from http:// www. thefreedictionary.com/helplessness

Wallach, L. & Wallach, M. A. (1999). Why is experimentation in psychology often senseless? *Scandinavian Journal of Psychology* 40: 103–106.

Wallach, M. A. & Wallach, L. (1998a). Of surrogacy, circularity, causality and near-tautologies: a response. *Theory and Psychology* 8: 213–217.

Wallach, M. A. & Wallach, L. (1998b). When experiments serve little purpose: misguided research in mainstream psychology. *Theory and Psychology* 8: 183–194.