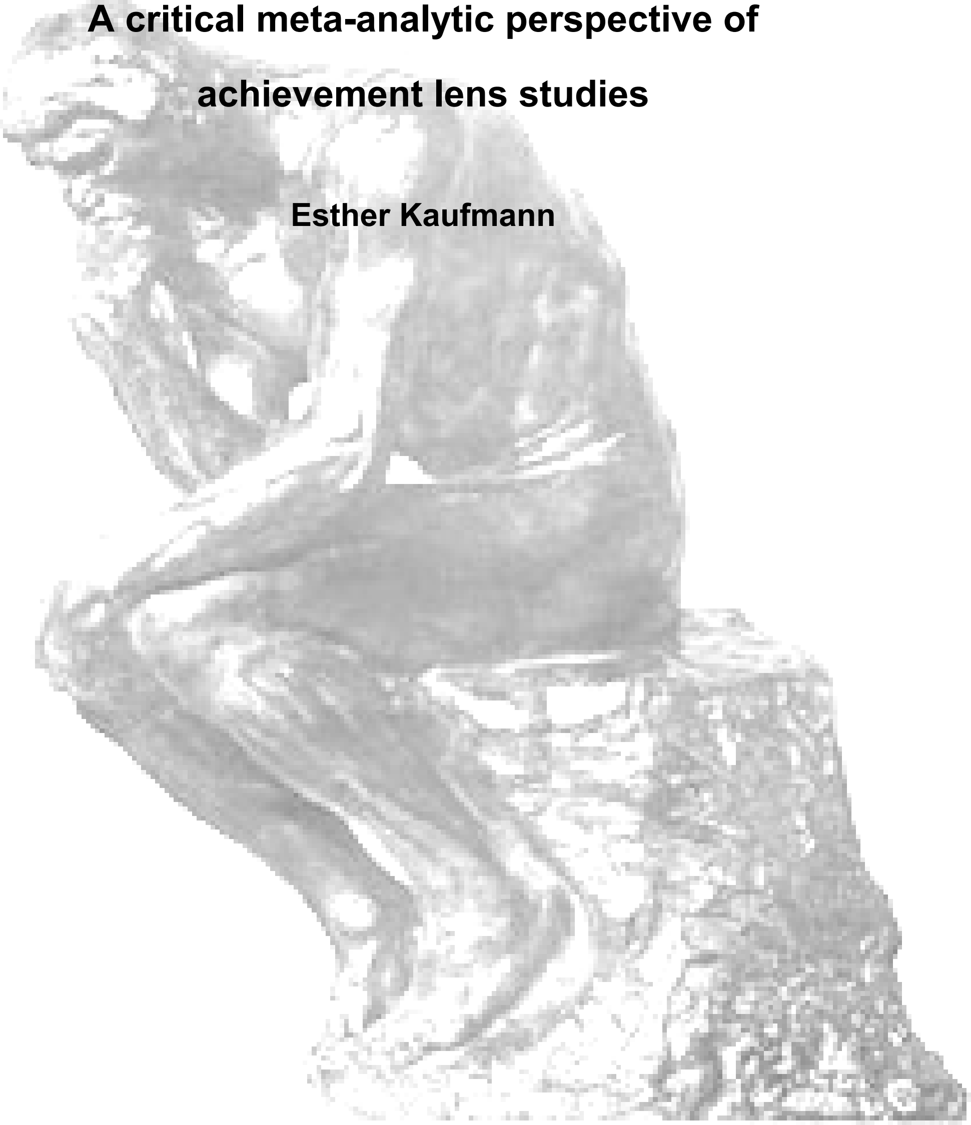
Flesh on the bones:

A critical meta-analytic perspective of achievement lens studies

Esther Kaufmann



Dissertation thesis written at the Center for Doctoral Studies in the Social and Behavioral Sciences of the Graduate School of Economic and Social Sciences and submitted for the degree of Doctor of Philosophy (Ph.D.) of the Faculty of Social Sciences at the University of Mannheim.

Academic Director: Prof. Dr. B. Ebbinghaus

Supervisor: Prof. Dr. W. W. Wittmann

Co-Supervisor: Prof. Dr. U.-D. Reips

Defense: 30. September, 2009

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to a number of people and institutions, without whose help this work would not have been possible.

First of all, I would like to thank my supervisors, Prof. W. W. Wittmann and Prof. Dr. Reips as well as Dr. J. A. Athanasou and Dr. L. Sjödahl for their enormous knowledge and great humanity that influenced me profoundly.

Secondly, it was a great honour and pleasure for us that our project was also supported by the Brunswik Society, namely, Prof. Hammond and Prof. Wolf. My deep thanks also to Prof. Wilkening, Prof. Scholz, Prof. Jonas, and Dr. Mutz for their advice and support. Dr. Karelaia and Prof. Hogarth for their meta-analyses, which supplement ours.

Thirdly, I would like to thank the Graduate School for providing me with its infrastructure. For feedback to our work, which was highly appreciated: Prof. Geschwend, Prof. Erdfelder, the CDSS students, Salina Yong, Gillian Sjödahl, Dr. Waldkirch.

I would like to acknowledge the authors of the studies used in our meta-analysis, without whose work it wouldn't have been possible to realize such an interesting project.

Taken together, this work gave me the opportunity to profit from enormous expert knowledge and to live abroad in Mannheim. I'm enormously grateful for this experience.

Beside the academic field, I would like to thank my parents, Elisabeth and Paul Kaufmann, Barbara Brettschneider, and my sisters Madlen Kaufmann, Gaby and Patrick Steiner for their understanding and support. Finally, without Phil Wyniger I would miss something special, thank you.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF TABLES

Page

APPENDICES

Appendix B: Literature search

Appendix D: Comparison with the meta-analysis by Karelaia and Hogarth (2008)

Appendix E: Psychometric meta-analysis according to Hunter and Schmidt (2004)

Appendix F: Results of our idiographic-based meta-analysis

Appendix G: Results of our nomothetic-based meta-analysis

Appendix H: Results of our robustness analysis

Appendix I: Bias-adjusted R^2

LIST OF FIGURES

Page

APPENDICES

Appendix F: Results of our idiographic-based meta-analysis

Appendix I: Bias-adjusted R^2

Appendix J: Success of single expert models

LIST OF EQUATIONS

Page

APPENDICES

Appendix C: *LME* component calculation

Appendix E: Psychometric meta-analysis according to Hunter and Schmidt (2004)

ABSTRACT

The major purpose of probabilistic functionalism is to appraise the “… interplay and relative contribution of environmental factors in the (organism’s) adjustment to a given ecology” (Brunswik, 1956, p. 143), the Lens Model Equation is of utmost importance because it permits the precise analysis of the “interplay” (Hammond, 1966, p. 72), well known as judgment achievement. Consequently, our meta-analysis on the Lens Model Equation leads to evaluation of the mind adaptation assumption in five different research areas. To prove this we used idiographic and nomothetic data of Lens Model Equation studies to prevent any fallacy (ecological vs. individualistic). In our analysis regarding the experience level within areas only business students’ judgment achievement indicated moderator variables. In all areas except in psychology judgment achievement is almost moderate. In addition, in our psychometric analysis judgment achievement clearly increases, but the values in psychology science are still low. Different sensitivity analysis supported the robustness of our results that imply area differences in experts judgment achievement.

1 INTRODUCTION

Important decisions, for example whether or not to get married, and to whom, we make rarely. Decisions such as which shoes to wear in the morning or at what time we actually get up, we make daily. Furthermore, some of us get paid to make correct decisions as experts, like physicians or teachers. Teachers, for instance, estimate the reading abilities of students, which influences their further school career. Consequently, our judgments and also those of experts greatly affect our personal and public life. Therefore, the question arises, how good the experts' judgment, on which our public life depends, actually is. Or, why are some people more accurate in their judgments than others? Better quality of judgment is badly needed in areas such as education and medicine, where it could improve human conditions and save many lives.

The ultimate goal of judgment and decision-making research is to improve a person's judgment. In the research carried out on judgment and decision making, differential judgment achievement is the central issue. By reviewing judgment achievement and the underlying cognitive processes, it is possible to make recommendations for improving a person's judgment and decision making.

In the following, an array of approaches to judgment and decision making is overviewed by the Cognitive Continuum Theory (*CCT*, Hammond, 2007). The *CCT* shows that most research uses an experimental design. Studies with an experimental design based on variance analysis, as for example the Heuristic and Bias school of thought (Kahneman, Slovic, & Tversky, 1982). This approach, as many others, implies a bad picture of human decision makers as mostly biased. In order to evaluate whether humans are really bad decision makers we should urgently supplement the experimental approach (or variance analysis approach) point of view with a correlative approach like the Social Judgment Theory (*SJT*).

The *SJT* is (Brehmer, 1988):

> A general framework for the study of human judgment. Despite its name, it is not a theory for it provides no testable hypotheses about judgment. Instead, it is a meta theory, which gives direction to research on judgment. (p. 13)

Furthermore, the focus on the *SJT* allows us to consider the validity (APA, 1954) and the aggregation problem (Robinson, 1950; Wittmann, 1985) in judgment and decision-making research, which is mostly neglected by other judgment and decision-making approaches.

This dissertation considers how judgment achievement across and between persons varies in the framework of the *SJT,* where it is defined as “the degree of correlation between a judge’s responses to cue profiles *…* and the criterion measurements for those profiles” (Cooksey, 1996, p. 367). For example, it is the degree of correlation between 1) *judgments*, such as a teacher’s reading-achievement estimation based on *cues* like socio-economic status of the students, and 2) the *criterion,* such as the end-of-year reading achievement of students measured by a test. Judgment achievement can be described by components of correlations, called the Lens Model Equation (*LME,* Tucker, 1964). Studies applying the *LME* include judgment tasks in which the three factors mentioned above are known. As a result, the components of the *LME* applied to judgment tasks can show how judgments come about.

The *LME* has been applied in numerous contexts to individuals (i.e. idiographic approach) or across individuals (i.e. nomothetic approach).

However, no comprehensive meta-analysis of the *LME* has been published that includes also individual data*.* But this is needed to check the ecological fallacy (Robinson, 1950). In addition, we used a meta-analytic approach according to Hunter and Schmidt (2004) to overcome the individualistic fallacy. This meta-analytic approach was selected, since the estimated population correlation can be corrected by the observed correlation for downward bias due to various artefacts, such as measurement error. In comparison to other meta-analytical approaches, the Hunter and Schmidt method also offered the best estimate of the population parameter (see Field, 2001, 2005). Hence, conducting a meta-analysis according to Hunter and Schmidt (2004) is also to determine whether the variance in reported components of *LME* was entirely the result of artefacts like sampling or measurement error. Therefore, this psychometric meta-analysis approach according to Hunter and Schmidt (2004) describes judgment achievement in the framework of the *SJT* in order to determine the actual judgment achievement of individuals or across individuals.

Furthermore, to find out why some people are more accurate than others, we meta-analysed the judgment achievement over all studies using an idiographic research approach with the components of the *LME*. Firstly, the error-free judgment achievement (*knowledge*) component describes the correlation between *judgments* and *criterion*, assuming the judge is perfectly consistent and the environment is perfectly predictable. Secondly, the consistency component reveals how perfectly consistent a judge actually is, expressed as a correlation between *cues* and *judgments*. Finally, the environmental predictability is expressed as the correlation between the *cues* and the *criterions*.

In addition, the meta-analysis was repeated with studies also using a nomothethic research approach and checked for possible moderator variables (i.e. applied research area, experience level).

Finally, the goal of this dissertation is – by means of a psychometric meta-analysis according to Hunter and Schmidt (2004) – to permit a first-time overview of judgment achievement across and between different studies in the framework of the *SJT*. This is urgently needed to evaluate the *SJT* approach and to relate the results to other judgment and decision-making theories. This evaluation of judgment achievement according to Hammond's *CCT* (2007) is more precisely described in the following chapters.

2 THEORIES ON JUDGMENT AND DECISION MAKING

A major concern in psychology is to understand judgment and decision making (*JDM*). The field of *JDM* has developed over the last 50 years and is an important precursor of modern cognitive psychology. During this time, psychologists have proposed various approaches to researching *JDM*. Brehmer (1987) summarized critical points of *JDM* research as follows:

> Psychological research does not provide any unified picture of human judgment either there is a variety of theoretical approaches to judgment, each with its own definition of the term. (p. 199)

This heterogeneity in approaches and definitions is also represented in the different classification systems. For instance, there are reviews separating correspondence vs. coherence theories (see Hammond, 2007), or normative, descriptive, and prescriptive decision theories (see Baron, 2004; Scholz, Mieg & Weber, 2003). To simplify the following overview, we will focus on only one classification system, namely Shanteau's (2001, see Figure 1 for an overview) classification of normative and descriptive theories. As the modern history of research on *JDM* has been dominated and started with the normative theory, we will present this theory first. Then, we will introduce the division of normative theories into riskless and risky judgments with an example. Second, the descriptive theories will also be presented. Finally, the Social Judgment Theory (*SJT*), combining both theories, will be explained in detail.

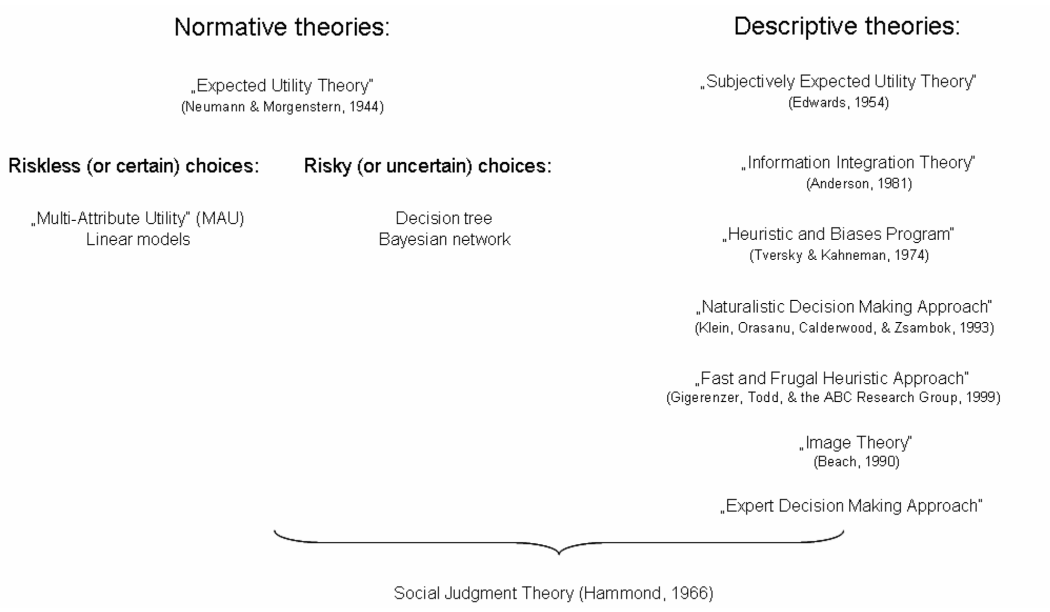

Normative theories:
„Expected Utility Theory"
(Neumann & Morgenstern, 1944)
Riskless (or certain) choices:
„Multi-Attribute Utility" (MAU)
Linear models
Risky (or uncertain) choices:
Decision tree
Bayesian network
Descriptive theories:
„Subjectively Expected Utility Theory"
(Edwards, 1954)
„Information Integration Theory"
(Anderson, 1981)
„Heuristic and Biases Program"
(Tversky & Kahneman, 1974)
„Naturalistic Decision Making Approach"
(Klein, Orasanu, Calderwood, & Zsambok, 1993)
„Fast and Frugal Heuristic Approach"
(Gigerenzer, Todd, & the ABC Research Group, 1999)
„Image Theory"
(Beach, 1990)
„Expert Decision Making Approach"
Social Judgment Theory (Hammond, 1966)

Figure 1. Classification of decision theories according to Shanteau (2001).

2.1 Normative theories

According to Over (2004), "normative theories tell us how we should ideally make judgments and take decisions" (p. 4). Normative theories are concerned with the development and application of models based on formal logic derived from economics (e.g. Expected Utility Theory) or statistics (e.g. Probability Theory). Researchers take these models as norms. These norms are standard or "benchmark", against which judgments are evaluated. On the one side, if judgments systematically deviate from the proposed models, this is called bias. On the other side, if both – judgments and "benchmark" – match, this implies that the judgment is correctly or optimally made. Hence, in this approach, the models are the golden standard to be reached by good judgment. The suggested golden standard fully describes how people would behave if they followed certain requirements of rational decision making. According to this, the decision maker is like a "rational actor". Hence, any rational person would follow the proposed models.

Historically, the impetus for the normative approaches to research on *JDM* comes from the seminal book *Theory of Games and Economic*

Behavior (Neumann & Morgenstern, 1944, see Figure 1), introducing the classical "*Expected Utility Theory*".

In addition, as you can see in Figure 1, Shanteau (2001, p. 55) divided normative theories into riskless (or certain) choices and risky (or uncertain) choices. These are explained in more detail in the following.

2.1.1 Riskless vs. risk judgments

Although Shanteau (2001) introduced two riskless judgment theories, the Multi-Attribute Utility Theory (see Edwards & Newman, 1982) and the linear models, we only focus on linear models because of its relevance to our meta-analysis.

Linear models (e.g. a regression model, see Dawes & Corrigan, 1974) have been used to describe judgment under certainty. For example, Dawes (1971) applied linear models to the selective admission of psychology graduate students at the University of Oregon. His research showed that a linear model of three quantitative admission variables (graduate record exam score, grade point average, and a crude index of the quality of the undergraduate institution) was consistently the best predictor of success in graduate education. This robustness of linear models – well-known as the beauty of linear models (Dawes, 1979, see also chapter 2.4.1.2) – is also confirmed by many other studies.

In addition, to describe risk judgments, Sheanteau (2001) introduce two further normative models; the *decision tree* and the *Bayesian network* (see Figure 1). As both theories are not relevant for our meta-analysis we refer to Shanteau (2001) for an overview.

2.2 Descriptive theories

In comparison to normative theories, descriptive theories have mostly been used. “Descriptive theories in psychology try to describe how people actually think” (Over, 2004, p. 4). To introduce descriptive theories, Sheanteau highlighted the *Subjectively Expected Utility Theory*, the *Information Integration Theory* (Anderson, 1981), the *Heuristic and Biases Program* (Tversky & Kahneman, 1974), the *Naturalistic Decision Making Approach* (Klein, Orasanu, Calderwood, & Zsambok, 1993), the *Fast and Frugal Heuristic Approach* (Gigerenzer, Todd, & the ABC Research Group, 1999), the *Image Theory* (Beach, 1990, see Figure 1). However, as a consequence of the mentioned research on *JDM*, the need for psychologists to help professionals make better decisions was recognized and encountered with the *Expert Decision Making Approach*; for an overview see Shanteau and Stewart (1992). To add that is this approach supports also our analysis (see chapter 2.5.2.1).

Finally, the *SJT* was also included in descriptive theories on *JDM* research, although this approach combines the normative and descriptive theories (see Shanteau, 2001, p. 554). In the following section, we will introduce the *SJT* and focus on the reasons for its selection for our meta-analysis.

2.3 Criticism of judgment and decision-making research

Despite the differences between normative and descriptive approaches, there have been many successful applications of both theories on *JDM* in many settings. However, the interesting question, whether decision makers are fully rational or biased, or simply how accurate judgment and decision makers are, is still unanswered. In addition, two major critique points are described. First, the validity problem, or the required increase of external validity on *JDM* research. Second, the nomothetically orientated research on *JDM* without considering the aggregation problem, and the resulting neglected focus on idiographic research approach.

2.3.1 Are decision makers biased?

In the following, we will illustrate the answer to the question whether decision makers are biased in a chronological order.

In the introduced normative approach, researchers do not insist that people never make mistakes in their judgments, but they do insist that the mistakes are unsystematic. The deviations of the judgments from an objective value (optimizing model) are called biases. Additionally, as previously noted, research shows that the principles of normative theories are systematically violated by the decision makers. Edwards (1968) and his colleagues, for instance, concluded that human judgment does not accord with a model of Bayes' rule for making judgments.

Subsequent research showed that the judgment maker is not fully rational as implied by the normative theories. Therefore, a new theoretical approach was developed. From the perspective of the new approach, the optimizing model was an unrealistic standard for human judgment. This standard excludes that the world is large and complex and we do not have the capacity to understand everything. We also have a limited time in which to make decisions. Therefore, Simon (1955, 1956) proposed a more limited criterion and introduced the concept of "bounded rationality" in decision making. According to Simon, people are fully rational, but only if they are not restricted by task aspects, such as time limit, or personal aspects like computational capacities. Hence, decision makers would make rational judgments, if they could gather and process sufficient information.

The concept of "bounded rationality" was not in line with the view of the Heuristic and Biases Approach. Tversky and Kahneman (1974) conducted experiments, in which the conditions are optimal; this means that the persons could gather and process sufficient information for their judgments. However, the Heuristic and Bias Approach implied a more negative view of human judgment making compared to the bounded rationality approach. In addition, in the time to follow, people were trained to overcome or avoid errors of judgment. But these efforts were largely deemed to be unsuccessful. Consequently, the studies generally lead to the conclusion that “things are even worse than we thought; not only is judgment incompetent, it resists remedial efforts” (see Hammond, 1996, p. 204). The Heuristic and Biases Approach is also highly criticized, however (see Hammond, 1996, p. 204).

In sum, there is a large body of findings accumulated from research on *JDM*. The reported modern history of research starts with an optimal view of decision makers and ends with the contrary: A view of decision makers as almost always biased, as implied by the Heuristics and Biases Approach.

However, to reveal the weakness of the conclusion that decision makers are often biased, we are interested to relate this statement to a comprehensive overview of the different applied decision theories, which are clarified in the following.

2.3.2 Cognitive Continuum Theory

To give a comprehensive overview of judgment and decision-making research, the Cognitive Continuum Theory (*CCT*, see Hammond, 2007) is applied to the described decision theories. Cooksey (1996) introduced the *CCT* as follows:

> Hammond proposed the *CCT* as a unifying theory for the field of human judgment and decision making. He intended to integrate, not replace, the currently popular, yet disparate, theories in the field. (p. 13)

Furthermore, the focus of the *CCT* is the relation between the judge and the task. Task properties are considered important, because they influence judgments. The *CCT* groups task properties into three principal categories: 1) complexity of task structure, 2) ambiguity of task content, and 3) form of task presentation (e.g. number of cues, reliability of cues, interrelationships among cues). The different task structures call for a kind of thinking on a continuum from analytic to intuitive thinking. Consequently, there are three continua on which any judgment will fall. According to Hammond (2007), “one must keep these [three dimensions] in mind when trying to understand performance in any judgment situation” (p. 129). A good illustration of this is the study by Hammond, Hamm and Grassia (1986).

However, because we are interested in giving a comprehensive overview on *JDM* research, we refer to Hammond (2007, p. 123; Cooksey, 1996, p.13) for detailed information about the *CCT* and show it in Figure 2. As you can see in this Figure, the described *JDM* theories can be embedded in it and are separated into coherence and correspondence approaches (see Cooksey, 1996):

> A coherence-based focus describes, explains, or predicts judgment competence on the basis of logical, mathematical, or statistical rationality. The interest is in whether or not the judgment is consistent with what some well-established set of rules or axioms would have produced, not the accuracy of a judgment with respect to some environmental criterion. A correspondence-based focus, on the other hand, describes, explains, or predicts judgment competence on the basis of its empirical accuracy. Here, the interest is in how well judgments map onto events in the world, not in the fact that judgments may have followed some internally consistent set of rules or axioms. (p. 44)

To summarise: Boxes two and three represent all normative and some descriptive theories representing coherence theories. As a consequence, to get an overview on humans’ abilities of *JDM,* more research on the side of correspondence theories is needed, such as the *SJT*.

To complement the resulting conclusion that more research with a correspondence-theory approach should be carried out, two critique points – the *validity problem* and the *aggregation problem* – in judgment and decision-making research are considered in more detail in the following. Both critique points support the application of the *SJT* as well.

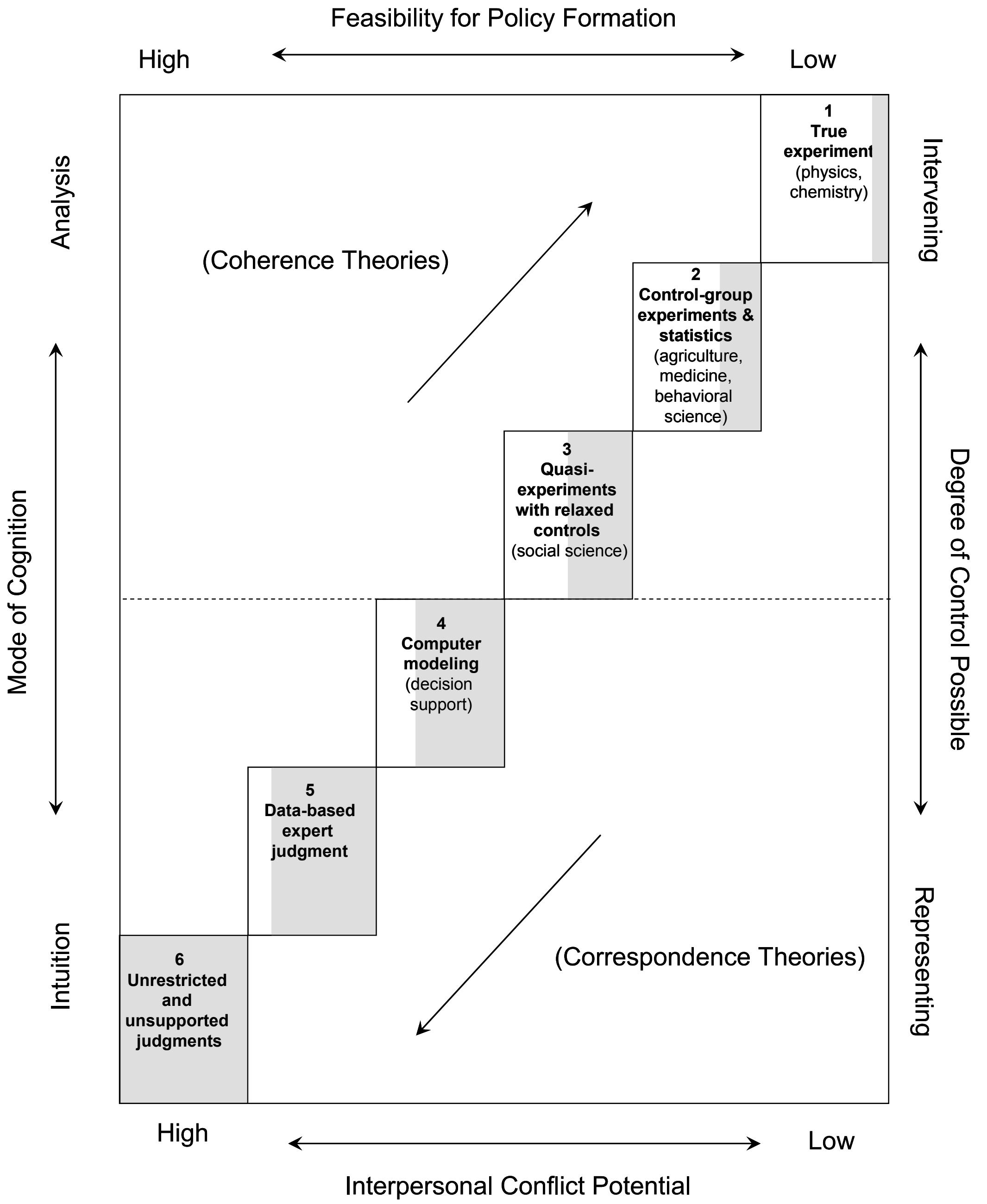

Feasibility for Policy Formation
High
Low
1
True experiment
(physics, chemistry)
Intervening
Analysis
(Coherence Theories)
2
Control-group experiments & statistics
(agriculture, medicine, behavioral science)
3
Quasi-experiments with relaxed controls
(social science)
Mode of Cognition
Degree of Control Possible
4
Computer modeling
(decision support)
5
Data-based expert judgment
6
Unrestricted and unsupported judgments
Intuition
(Correspondence Theories)
Representing
High
Low
Interpersonal Conflict Potential

Figure 2. The *CCT* (modified from Hammond, 1996, p. 235)

2.3.3 Validity problem

Since the APA publication (1954) on validity, validity – the generalisation of psychological measurement – is also considered in research on *JDM*. Hogarth (according to Hammond, 2007) described the goal of *JDM* research to generalize the results as follows:

> Researchers who study people's decision making processes seek results that are generalizable. However, conclusions are often based on contrived experimental 'incidents' with little understanding as to how these samples of behaviour relate to the population of situations that people encounter in their naturally occurring environments (i.e., the so-called real world). (p. 217)

Most *JDM* researchers used experimental designs, i.e. controlled conditions of the laboratory. Cooksey (1996) critiques experimental research as follows: "organism behaves in under atypical conditions in pursuit of tasks and goals which were not representative of the natural environment in which the organism was embedded" (p. 1). Consequently, the studies are not *external* or *ecologically valid*, do not represent the real decision-making situation, and it is therefore also difficult to generalise the results of the studies. To raise the ecological validity of psychological science, Brunswik recommended that psychology should be a science of organism-environment relations rather than a science of the organism (see Dunwoody, 2006). This suggestion also influenced Newell and Simon (1972):

> Just as a scissors cannot cut paper without two blades, a theory of thinking and problem solving cannot predict behaviour unless it encompasses both an analysis of the structure of the task environments and an examination of the limits of rational adaptation to task requirements. (p. 55)

However, as Dunwoody (2006) noted:

> Brunswik's argument still carries weight today, and psychology in general, and cognitive psychology specifically, have still not dealt

> with the criticism levied against it by Brunswik 50 years ago. (p. 139)

Furthermore, as mentioned before within the *CCT*, theories are separated into coherence and correspondence approaches (see Cooksey, 1996). This classification of decision theories reveals that despite the importance of the characteristics of the task to raise the ecological validity in *JDM* research, most studies only concentrate on the coherence theories, and therefore on the characteristics of the judge. However, explanations of judgment are found both in characteristics of the judges, such as experience, and in characteristics of the task.

In sum: To find out, whether a decision maker is actually biased as implied by the Heuristic and Biases Program, it is also necessary to raise the external validity of the research on *JDM* and therefore to include more studies with a correspondence approach like the *SJT*.

2.3.4 Neglected idiographic approach

An important neglected critique regarding research on *JDM* concerns the fact that most studies use the nomothetic approach.

The historical background leads to the introduction of the terms which resemble two disciplines of sciences. At the beginning of the last century, different disciplines dominated science, and each of them wanted to influence the others with its methods. To avoid methodological confusion, Windelband suggested two disciplines, as nomothetic disciplines (e.g. natural science) seek only general law – in contrast to idiographic disciplines (e.g. history), which seek to understand a particular event. Windelband's (1894) definition of "nomothetic" (greek: "nomos" = law) and "idiographic" (greek: "idios" = own, private) science as two distinct kinds of knowledge is traceable to Aristotle and to Kant.[1]

> Nomothetic knowledge, Windelband argues, is knowledge of the sort contained in the general laws formulated in the natural

[1] Windelband was a member of the Southwestern School of Neo-Kantianism.

> sciences (Naturwissenschaften). The defining characteristic of a general law is that it reflects “what always is” within some explicitly circumscribed domain of empirical events covered by the law.
>
> Idiographic knowledge by contrast, is knowledge of an essentially historical or biographical sort. Its defining characteristic is its reflection of “what once was”, and so idiographic knowledge is precisely that sort of knowledge needed to understand some unique entity or event. … sought in the Geisteswissenschaften, or what would be referred to English as the moral sciences or human sciences, or, most commonly, the humanities (see Lamiell, 2003, p. 89).

Later, as mostly reported, Allport (1937) introduced Windelband’s distinction between the idiographic and the nomothetic approach to psychology. However, Hurlburt and Knapp (2006) argue that it is often overlooked that the terms were already part of the psychological discourse of the “leading logician”, Hugo Münsterberg, in 1898. In turn, Münsterberg had a strong influence on Stern, who is the pioneer of individual psychology. This new psychology had its focus on a new unit, the person as the “unitas multiplex” (see Kreppner, 1992, p. 539). Stern also influenced Allport, who collected terms describing personal characteristics and found that some of them could be investigated at a nomothetic level. But, the majority of these terms are more or less unique dispositions based on life experiences, and they introduced an idiographic level of research in Psychology. Therefore, Allport (1937) argued that the psychology of personality needs both approaches as follows:

> The dichotomy [between nomothetic and idiographic], however, is too sharp: It requires a psychology divided against itself… . It is more helpful to regard the two methods as overlapping and as contributing to one another. In the field of medicine, diagnosis and therapy are idiographic procedures, but both rest intimately upon knowledge of the common factors in disease determined by the nomothetic sciences of bacteriology and biochemistry. Likewise,

> biography is clearly idiographic, and yet in the best biographies one finds an artful blend of generalization with individual portraiture. A complete study of the individual will embrace both approaches. (p. 29)

In summary, Stern and Allport's research experience leads to the introduction of idiographic and nomothetic approaches also in psychology.

In the following, many scientists argue that the idiographic approach does not belong in the realm of science, because the focus of science is on the development of universal, nomothetic laws of behavior. Therefore, it would be impossible to generalize results (for a review of such critiques, see Runyan, 1983).

Nevertheless, of the criticism mentioned above, some scientists' also focus on the idiographic approach (see Asendorpf, 2000; Molenaar, 2004). For example Brunswik (1956) concerned the uniqueness of each organism, as it engaged in functional behaviour within the context of a particular ecology, and developed his probabilistic approach. This approach is used by Hammond (1955) to study cognitive processes. The study of cognitive processes on the individual level is also supported by Newell and Simon. In line with Brunswik, they prefer the analysis of the cognitive activity of each individual separately (see Newell & Simon, 1972, p. 874).

Consequently, we can conclude that also dominant cognitive psychologists like Newell, Simon, and Hammond use individual data in the same way as Allport and those who preceded him, who emphasize the importance of the idiographic approach not as a substitute for or an enemy of the nomothetic approach, but as an informer and companion of it.

2.3.4.1 Aggregation problem: Ecological vs. individualistic fallacy

Today, psychology, and especially research on *JDM*, is dominated by the nomothetic research approach, which investigates large groups of people in order to find general laws of behavior to apply to everyone. With the nomothetic approach, however, the aggregation of individual data could produce misleading interpretations (Asendorpf, 2000), such as the ecological fallacy (Robinson, 1950). The ecological fallacy arises because associations between two variables at the group level (or ecological level) may differ from associations between analogous variables measured at the individual level. In his study, Robinson (1950) computed the literacy rate and the rate of the foreign-born population for each of the 48 states in the USA. The correlation between the 48 pairs of numbers was .5 – so, the greater the proportion of immigrants in a state, the higher its average literacy. This is an ecological correlation, because the unit of analysis is a group of people, not an individual. In contrast, on the individual level the correlation was lower (-.1), so immigrants were on average less literate than native citizens. The positive correlation at the level of state populations resulted because immigrants tended to settle in states where the native population was more literate. Therefore, the ecological correlation gives a wrong result, because using averages diminishes the variability in the underlying individual data. Thus, Robinson cautioned against drawing conclusions about individuals on the basis of the aggregation level, or “ecological” data. As Robinson was first to mention this problem in the aggregation of correlation data, this fallacy was also given the name Robinson effect.

Furthermore, the ecological fallacy is not only a problem in aggregation within one study, but also in aggregation across studies in meta-analysis research (Viechtbauer, 2007, p. 114). In meta-analysis, the unit of analysis is mostly studies, not the individual participant within a study, and consequently, meta-analysis summarizes relationships at study-level, and these relationships may not correspond to the observed relationships at the individual level.

In addition, also the individualistic fallacy (i.e. exception, psychological or atomistic fallacy) – the counterpart of the ecological fallacy – should be considered. The individualistic fallacy occurs when a group conclusion is reached on the basis of exceptional cases. An example could be a man cooking badly, from which we conclude that "men are terrible cooks". In short, a stereotype leads to our conclusion (for an overview of typology of ecological fallacies, see Alker, 1969).

To summarize: Although the aggregation of judgment achievement prevents individualistic fallacy, ecological fallacy should not be overlooked.

In the following, we emphasize that some modern cognitive psychologists are aware of the importance of the introduced aggregation problems. Cooksey, Freedbody and Davidson (1986, p. 49) mention the aggregation problem in their description of *JDM* research as follows: "Apart from the *SJT* work …., the study of judgment and decision making has tended to remain in the methodological traditions of nomothetic design and analysis" (p. 49). Furthermore, they contend that "an individual's decision system needs to be viewed in isolation and as a coherent whole, before aggregation across judges occurs" (p. 49). Consequently, they concluded that "aggregation [across individuals] can occur, but only after individual judgment policies have been completely specified" (p. 50). "Only after understanding the uniqueness of individual judgment policies will we be in a position to talk about the commonalities between policies – that is aggregation" (p. 50).

2.3.5 Summary of chapter 2.3

In sum, the critique on *JDM* – the validity and aggregation problems – and the *CCT* support the research on *JDM* in the framework of the *SJT*, which includes the Lens Model (Brunswik, 1952). The *SJT* is suitable to research *JDM*, because the Lens Model focuses on individuals and therefore on the idiographic approach. Additionally, the *SJT* includes also the nomothetic approach and, therefore combines both theories on *JDM*. At last, because the *SJT* also includes the correspondence theory, it can be shown how the judgment achievement of a person is influenced by the cognitive system of the judge and/or by task properties. Consequently, of all mentioned critique on research in judgment and decision making, we will focus on the *SJT* in more detail. As the *SJT* is embedded in Probabilistic Functionalism, it will be introduced first.

2.4 Probabilistic Functionalism

In the following we will introduce the background of the *SJT* in more detail. First, the Probabilistic Functionalism Approach, leading to the development of the *SJT*, will be introduced. Then, we focus on the classical Lens Model and the deduced Lens Model Equation (*LME*).

Probabilistic Functionalism is traceable to Brunswik's work on perception during the 1930s (see Wolf, 1995, p. 16 for biographical information on Brunswik). In those times the gestalt psychologists studied a wide variety of perceptual illusions and turned their attention to the study of error. Because Brunswik and also Gibson insisted that the study of illusions was misguided, both wanted to study behaviour in a natural environment, which lead them to become known as ecological psychologists. However, Brunswik's interest in the structure of the environment in relation with an organism is based on his early collaboration with Tolman (1935). In their publication "The Organism and the Causal Texture of the Environment", environment texture was already the focus. They assume that persons try to cope with an environment consisting of interrelated and thus "textured" objects and events. However, this point of view contrasts sharply with the theme psychologists took from physical science in those days and that led them to focus on the exact mathematical laws of behaviour (see Hammond, 2007). Beside this collaboration, Brunswik (1939) conducted his own rat experiments in Berkeley of Tolman's invitation. In these experiments his research showes that rats follow a probability-matching rule representing the probability of getting food or the environmental conditions.

Consequently, this research experience may have lead Brunswik to recognize the importance of the environment in its influence on cognition processes, the key feature for the development of Probabilistic Functionalism. Hence, the essence of Brunswik's theory and methodology is seen in his definition of psychology's task:

> As the analysis of the interrelationship between two systems in the process of "coming-to-terms" with one another, the assertion that psychology must treat each system with equal respect, the directive that psychology "… must also be concerned with the texture of the environment as it extends away from the common boundary" of the two systems (Hammond, 1966, p. 16)

In addition, Brunswik integrated theory and methodology into one unified system mostly applied to perception. Probabilistic Functionalism was the framework for this development. Probabilistic Functionalism introduced the idea of an environment that included probability relations among the variables of interest (e.g. reading ability and reading achievement) instead of perfect relations among the variables, the so-called determinism approach. In those days the deterministic approach was dominant.

The Probabilistic Functionalism Approach leads to an environment which is not perfectly predictable and thus uncertain from the viewpoint of an individual. Furthermore, Functionalism implies a "utilitarian, adjustment-centered biological conception of psychology which may be traced to Charles Darwin's view on the struggle for existence" (Brunswik, 1952, p. 55). Brunswik (1955) argues that the real world is an important consideration in research (see also Dunwoody, 2006). Every person lives within and interacts with an environment. Thus, psychological processes are adapted in a Darwinian sense to the environment in which they function.

In summary, according to the Probabilistic Functionalism Approach, our judgment is adapted to the environment in which the judged criterion is not perfectly related to information on which the judgment is based. Therefore, the criterion is not perfectly predictable for a decision maker.

In addition, the major purpose of Probabilistic Functionalism is to appraise the "… interplay and relative contribution of environmental factors in the (organism's) adjustment to a given ecology" (Brunswik, 1956, p. 143).

According to Goldstein (2004, p. 22), "Brunswik's Probabilistic Functionalism emphasizes 1) adjustment to the world, and 2) the mediated nature of that adjustment". Already Hursch, Hammond and Hursch (1964) noted the environment's importance to adjustment or judgment achievement as follows:

> Without a knowledge of the limitations placed on achievement by the statistical characteristics of a given ecological or response system, it is impossible (a) to evaluate a subject's achievement within that system, (b) to compare a subject's achievement across ecological situations which have different statistical characteristics, and (c) to understand why the subject's achievement was as high or low as it was. (p. 43)

It must be noted that Brunswik's conceptualization failed to take hold in 1940s and 1950s psychology. Since the 1960s, Brunswik's ideas have enjoyed a steadily increasing influence on research in a variety of areas, such as methodology (e.g. Hunter, Schmidt & Jackson, 1982; Wittmann, 1985), human-technology interaction (e.g. Kirlik, 2006), but most notably in the study of human judgment (e.g. Hammond, Hamm, & Grassa, 1986). Brunswik's influence on *JDM* starts with the application of the Lens Model to human judgments by Hammond (1955), as you can see below.

Finally, the importance of Probabilistic Functionalism is summarized by Hammond (2007):

> [Brunswik] demonstrated that the departure from the physical-science model of scientific work and the adoption of the biological

> model would allow us to solve the problem of how to harmoniously advance academic work and make more usable its results. (p. 265)

Therefore, Brunswik's Probabilistic Functionalism Approach is a biological rather than physical-science view of psychology.

2.4.1 Social Judgment Theory

In the following, the *SJT* within the Probabilistic Functionalism framework is outlined.

As previously noted, Hammond (1955) first transferred Brunswik's ideas from visual perception to social judgment as clinical psychodiagnosis. Hammond's work was inspired by Meehl's study (1954), which presented strong arguments in favour of substituting statistical prediction for clinical judgment in diagnostic tasks (see chapter 2.5.2). As a consequence of Hammond's work, a comprehensive perspective on *JDM*, called the *SJT*, was developed. *SJT* is, as Brehmer (1988) noted:

> Despite its name, it [*SJT*] is not a theory for it provides no testable hypotheses about judgment. Instead, it is a metatheory which gives direction to research on judgment. (p. 13)

The *SJT* evolved through the 1960s and 1970s, as Hammond and his colleagues synthesized research that applied the Lens Model (Brunswik, 1952) to judgments under the rubric of the *SJT* (Brehmer & Brehmer, 1988; Hammond, Stewart, Brehmer, & Steinmann, 1975).

The *SJT* is characterized by four varieties of the lens model (Dhami, Hertwig & Hoffrage, 2004, p. 964, for a pictoral representation of these designs, see also Hammond & Stewart, 2001, p. 472), namely the *single-system design*, the *double-system design,* the *triple-system design,* or *N-system design.* From this model, also a *hierarchical N-system model* is derived (see Cooksey, 1996).

In sum, the Lens Model is a useful framework for conceptualizing the judgment process for an individual judge or a group of judges.

2.4.1.1 Classical Lens Model

Beside the mentioned variants of the classical Lens Model, the *double-system design* is illustrated in more details in the following. Then we will focus on the deducted *LME*.

Cooksey (1996) defined the classical Lens Model – the double-system design – as follows:

> Brunswik's original conceptual device for depicting the fundamental unit of focus for psychology where the ecology and the person's cognitive system are accorded equal importance from a research perspective and the linkages between the two systems are made explicit. The lens analogy comes from the linkages between the various surface cues and the depth region of the ecology and the depth region of cognition being convergently focused like rays of light onto the distal criterion on the one side and onto judgments on the other. (p. 370)

The Lens Model is used to describe the judgment achievement of an individual judge. For example, as previously noted, the application of the Lens Model to the expectation of a student's year-long reading achievement (Cooksey et al., 1986) is illustrated in Figure 3.

In a typical lens-model study, a person (in our example, a teacher) considers a number of student profiles and makes an estimate of a criterion (e.g. y_e, end-of-year reading achievement) for each student. "A profile is a descriptive term, which refers to the configuration of cue values (*e.g. information such as the socio-economic status of students*) used to depict a particular case for judgment" [italics added] (Cooksey, 1996, p. 372, also called cue profiles, events, or cases).

In his Lens Model, Brunswik applied the key principle of parallel concepts. "This principle states, quite simply, that the ecological system and the cognitive system of the organism can and should be described using the same types of concepts" (see Cooksey, 1996, p. 3). Hence, the two parallel concepts, representing the left and the right side of the lens, are explained. The right-hand side of the lens model represents the teacher's estimates (Y_s) as a cognitive process (see also Figure 3, chapter 2.4.1). For each student, there is some objective criterion value (y_e, e.g. standardized test score) for his reading success at the end of the year. This is shown on the left-hand side of the model as the environment of the judgment task. The teacher's task is to make the "best" judgment possible of this criterion performance for each student. The judgment is based on available information (x_1, e.g. socio-economic status, reading ability) that is perceived to be related to or predictive of the end-of-year reading success. The available information is called cues. According to Cooksey (1996) cues are:

> Any numerical, verbal, graphical pictorial or other sensory information which is available to a judge for potential use in forming a judgment for a specific case and/or which can be available in the environment for making predictions about a criterion. (p. 368)

The cues are represented in the center of the Lens Model. In fact, different cues may be used at different times, i.e., they may substitute for each other; this is well known as "vicarious functioning". Although Heider rejects Brunswik's probabilism, he accepts the idea of the intersubstitutability of cues (see Hammond, 1996, p. 141). A good example of vicarious functioning is given by Hammond (1996) and shows that cues are redundant and thus intersubstitutable:

> When pigeons are unable to locate the sun because of a cloud cover, magnetic lines of force function vicariously for the sun. Thus, "under complete overcast, if the sun compass fails to operate, the second step seems to be achieved by a magnetic compass." (p. 115)

Vicarious functioning takes a key position within the Lens Model, as illustrated by Brunswik (1957):

> Vicarious functioning emcompasses both the divergent and the convergent part of the lens like patterns that characterize all achievement. In the field of cognition, it is the divergent part –

ecological validity – which is ecological and the convergent part – utilization – which is organismic. (p. 22)

Also Darwin recognized the parallels between vicarious mediation (of information in the environment) and vicarious functioning (in the organism, see Hammond, 1996, p. 162) which make-up the vicarious-functioning processes.

In sum, the ability to shift dependence from one cue to another, or vicarious functioning, is a great advantage in an uncertain environment that offers redundant information.

Furthermore, the vicarious functioning process underlies Brunswik's achievement concept. In our educational example, judgment achievement is the achievement of a teacher's expectations. In this context, the teacher's expectation is defined as how well a teacher can predict the real reading ability, expressed as a correlation between the teachers' reading-achievement estimate and the actual test score of a reading-achievement test (r_a, see Figure 3).

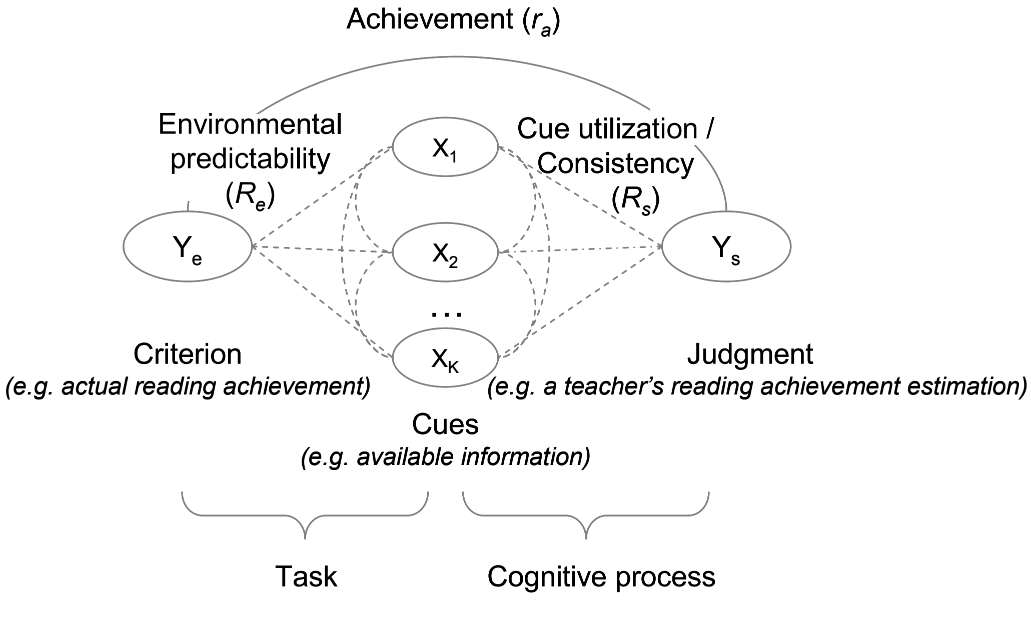

Achievement (r_a)
Environmental predictability (R_e)
Cue utilization / Consistency (R_s)
X_1
Y_e
X_2
Y_s
...
X_K
Criterion
(e.g. actual reading achievement)
Judgment
(e.g. a teacher's reading achievement estimation)
Cues
(e.g. available information)
Task
Cognitive process

Figure 3. Adapted Lens-Model representation for comparing a judgment with a known criterion (modified from Brunswik, 1952).

As described before, according to the Probabilistic Functionalism Theory, the cues are not perfectly related to the judged criterion in the environment. Therefore, in our example, even with optimal use of the information (cue utilization, consistency, R_s, see Figure 3), a teacher's reading-achievement estimate will not be perfectly accurate. Ecological validity or environmental predictability (R_e, see Figure 3) limit the judgment achievement. In the following, the term task predictability will be used, although the term ecological validity was introduced for it by Brunswik. For a discussion about the terminology we refer to Hammond (1998).

Achievement can be maximized, however, when the available information is strongly related to the actual-reading achievement test or when high validity of the available information exists. However, the environment is not deterministic, which leads to not prefect predictability of an event or judgment achievement.

2.4.1.2 Lens Model Equation

The achievement correlation (r_a), also known as validity index, can be decomposed into several components, combined in the so-called Lens Model Equation (*LME*). Consequently, limitations of judgment achievement are revealed by the components of the *LME*. One of the goals of the developers of the *LME* was to compare a subject's achievement across different situations.

The initiation for the development of the *LME* came in 1964 (see Hammond et al., 1964; Hursch et al., 1964; Tucker, 1964) and ultimately lead to the well-known Tucker *LME*. Tucker (1964) simplifies the *LME* by adding the component *G* (see Equation 1, and see Stewart, 2004, for biographical information about Tucker). Tucker's version of the *LME* became standard:

$$r_a = GR_sR_e + C\sqrt{1-R_e^2}\sqrt{1-R_s^2} \qquad (1)$$

To summarise: The *LME* consists of a linear (i.e. GR_SR_e) and a non-linear part. Mostly, the linear part explains the greatest part of judgment achievement. This also represents the introduced beauty of linear models (see Dawes, 1979, see chapter 2.1.1).

Since the early publication of the *LME*, several expansions of the basic *LME* have been developed. Castellan (1972), for instance, generalised the *LME* to multiple criteria, because in most judgment tasks we have to evaluate multiple criteria. He critized the classical Lens Model and the derivate *LME* as follows:

> The linear lens model as described is applicable only to situations involving a single criterion. But there are many judgment tasks involving several criteria. For example, in the case of clinical judgment, the decision concerning a patient may be whether the patient is schizophrenic and / or whether the patient should be hospitalized. (p. 244)

Hence, because the described Lens Model is applicable only to situations involving a single criterion (i.e. univariate Lens Model) and there are many judgment tasks involving several criteria, Castellan proposed a multivariate Lens Model. An example is the study by Cooksey et al. (1986). Another example of the extension of the *LME* is Stewart's (1976) hierarchical version. With this variation of the *LME,* it is possible to contribute different sets of variables. In addition, Castellan's and Stewart's expansion of the *LME* lead to the generalized *LME* by Cooksey and Freebody (1985). Finally, Stewart (1990, Stewart & Lusk, 1994) integrated Murphy's Skill Score into the Lens-Model concepts to give a more precise assessment of forecasting accuracy.

As the focus of our work is the classical *LME,* however, we will describe the application of the standard *LME* by Tucker (1964), derived from the Lens Model (see Figure 4). The *LME* reflects symmetry by the parallel application of two regressions, one to the organism and the other to the environment. Additionally, parallel components are derived for both systems expressed by the *LME*. Therefore, firstly, the regression is based on cues (*$x_{1\text{-}k}$*, independent variables) of multiple judgments, such as predictions of students' reading-achievement levels, by one teacher (Y_s, dependent variable). This first regression models the consistency component (R_s) of the *LME*. Secondly, in the same way, the environmental-predictability component (R_e) of the *LME* was modelled through the criterion value, such as the actual reading-achievement level of students (y_e, dependent variable) and the cues of multiple judgments.

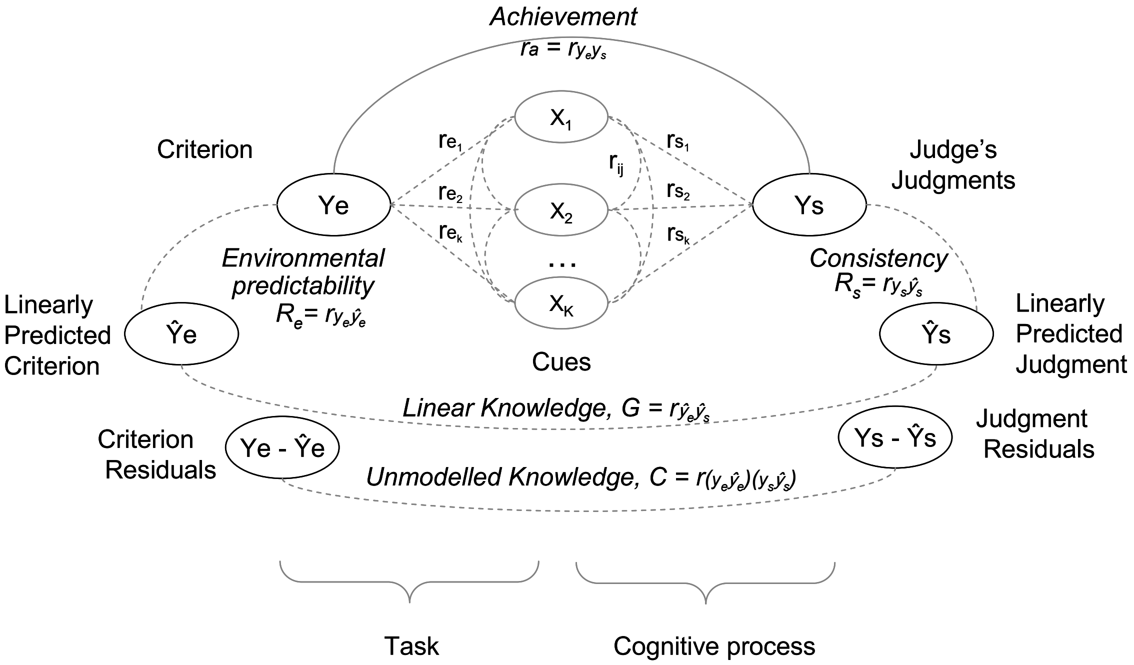

Achievement
$r_a = r y_e y_s$
Criterion
Ye
X_1
re_1
re_2
re_k
r_{ij}
X_2
...
X_K
rs_1
rs_2
rs_k
Ys
Judge's Judgments
Environmental predictability
$R_e = r y_e \hat{y}_e$
Consistency
$R_s = r y_s \hat{y}_s$
Linearly Predicted Criterion
Ŷe
Cues
Ŷs
Linearly Predicted Judgment
Linear Knowledge, $G = r \hat{y}_e \hat{y}_s$
Criterion Residuals
Ye - Ŷe
Ys - Ŷs
Judgment Residuals
Unmodelled Knowledge, $C = r(y_e \hat{y}_e)(y_s \hat{y}_s)$
Task
Cognitive process

Figure 4. Adapted Lens-Model with superimposed statistical parameters for comparing a judgment with a known criterion (modified from Cooksey, 1996, p. 206).

However, the results of the two mentioned regression analyses are correlated, leading to the judgment achievement. Hence, a teacher's prediction of reading-achievement level (*r_a*) can be described with the components of correlations of the *LME* as follows (see Table 1):

Table 1

Summary of the components of correlations of the LME (Tucker, 1964)

Component	Symbol	Description
Achievement	r_a	Correlation between the judgment and the criterion
Knowledge (error-free achievement, linear)	G	The correlation between the environmental predictability component and the consistency component
Consistency/ Cue utilization	R_s	The strength of the relation between the judgment and the cues (i.e. multiple correlation of the judgment and the cues)
Environmental / Task predictability	R_e	The strength of the relation between the criterion and the cues (i.e. multiple correlation of the criterion and the cues)
Knowledge (nonlinear)	C	The correlation between the variance not captured by the environmental predictability component or the consistency component (residual variance)

In the following, the single components of the *LME* illustrated by our educational example.

How well a teacher knows the reading-achievement criteria of a reading-achievement test is revealed by the components *G (linear knowledge)* and *C (nonlinear knowledge).* Therefore, the component *G (and also C*) can be considered to be an estimate of the correlation a teacher can achieve if the environment is fully predictable (R_e = 1) and if the teacher makes perfectly consistent estimates (R_s = 1). *G* represents the error-free judgment achievement of the judge and is called knowledge.

How similar repeated reading-achievement estimations of a teacher are is represented by the component R_s and is called consistency.

How well available information represents the reading-achievement criteria of a reading-achievement test is described by the component R_e. Therefore, the component R_e is called environmental predictability. For example, even if you are the best teacher in the world, you may be unable to predict the reading achievement of your students, because most of the variance in achievement comes form sources other than the used cues.

To summarize, in a more general way, the *LME* is a precise, mathematical way of describing the judgment achievement (r_a) of a person by four components (G, R_s, R_e, C). Furthermore, the *LME* is used to identify the underlying sources of judgment achievement. This equation is of utmost importance, because it permits the precise analysis of the interplay and relative contribution of environmental factors in the (organism's) adjustment to a given ecology.

The success of the regression-based *LME* in current research was not predictable. On the one hand, in the early days, there were strong critics against this approach. Hildegard (1955) summarized the critique against Brunswik's approach as follows: "Correlation is an instrument of the devil" (p. 228). On the other side, this approach coined the metaphor "man as intuitive statistician", as regression analysis is used to model how the mind works. However, the numerous publications in the framework of the *SJT* show clearly that correlation-based research is today widely accepted by the scientific community and leads to reviews on this research. In the following, we will introduce these reviews on *SJT*-based research.

2.5 Reviews on judgment achievement

In the following, we separate achievement from feedback and learning studies within the *SJT* (see chapter 4.1.2). Then, we will introduce research on *LME* components, leading also to presentation of the symmetry concept. Finally, to prove a complete overview of judgment-achievement studies, we will relate this *SJT* research to other *JDM* approaches – focusing on research areas and expertise knowledge, and leading to our research questions.

2.5.1 Within the Social Judgment Theory

A comprehensive overview of the area of the *SJT* is the "Role of Representative Design in an Ecological Approach to Cognition" (Dhami et al., 2004). In line with Dhami et al. (2004, p. 964), the following overview of the *SJT* is separated into research areas like a) judgment achievement, b) multiple-cue probability learning, or c) cognitive-feedback studies applying the lens model.

Firstly, of all *multiple-cue probability learning* studies until 1999, there is an annotated bibliography available by Holzworth, including 315 references. The subject of multiple-cue probability learning studies is how or how well an individual learns probabilistic (not perfectly linked) relations between two variables (e.g. reading ability and reading achievement).

Secondly, a complete literature review on *cognitive-feedback* studies until 1989 was done by Balzer, Doherty and O'Connor. These studies involve periodic information about the subjects' judgment strategies. Therefore, cognitive feedback may include a summary measure of past performance and/or information about the association between each cue and the subject's judgment. Furthermore, in their review Balzer et al. (1989) make explicit three types of feedback, namely: information about task, cognition, and functional validity. In summary, the data showed that many studies suggest that most of the benefit comes from the task information.

Finally, although *judgment achievement* should be the main research topic in psychology (Brunswik, 1966), it is often neglected in current research (see Dunwoody, 2006), as Dhami et al. (2004) point out:

> The majority of studies outside the Brunswikian tradition (97%) and the neo-Brunswikian studies (72%) described participants' judgment policies and compared policies among participants without reference to their degree of achievement. (p. 967)

However, there are numerous publications also on achievement studies leading to two meta-analyses (Karelaia & Hogarth, 2008; Stewart, 1997). We will present them in a historical order and relate them to our research.

2.5.1.1 Meta-analysis by Stewart (1997)

More than ten years ago, Stewart (1997) carried out a meta-analysis of experts' judgment achievement and ecological validity. In this meta-analysis, only two components of five lens-model components were considered, and they focus only on experts. Therefore, our meta-analysis should be extended to all components of the *LME* and include non-experts' judgments. To mention is that this meta-analysis was not available, hence, we can't compare our results with it.

2.5.1.2 Meta-analysis by Karelaia and Hogarth (2008)

During the work on this dissertation-thesis we became aware that Karelaia and Hogarth (2008) also carried out a meta-analysis in the framework of the *SJT*. However, there are four differences in these two meta-analyses, which are explained in the following:

1) Karelaia and Hogarth's meta-analysis includes only a *bare-bones meta-analysis*. According to Hunter and Schmidt (2004) and in contrast to Karelaia and Hogarth (2008), we include also measurement-error corrections, because:

> A theory of data that fails to recognize measurement error will lead to methods of meta-analysis that do not correct for measurement error. Such methods will then perforce produce biased meta-analysis results. (p. 31)

Hence, Karelaia and Hogarth's meta-analysis implicitly assumes perfect reliability or the absence of measurement error, which is clearly not the case in any study in science. Furthermore, Hunter and Schmidt's research experience (2004, p. 68) showed that when they made corrections for sampling error and other artefacts, they usually found little appreciable variation across studies remaining after these corrections. Finally, they also wished to inform that researchers should not forget that even a fully corrected meta-analysis suggested by their 2004 book will not correct for all artefacts. And therefore, they conclude: "Even after correction, the remaining variation across studies should be viewed with scepticism. Small residual variance is probably due to uncorrected artefacts rather than to real moderator variable" (p. 81).

2) In addition to Karelaia and Hogarth's meta-analysis, we will also focus on studies with an idiographic research approach and can therefore check our data for the ecological fallacy (see Robinson, 1950).

3) Furthermore, Karelaia and Hogarth's data base, including *cognitive feedback studies* and also *multiple-cue probability learning studies*, is different from our studies and complements this dissertation-thesis.

4) Finally, Karelaia and Hogarth (2008) neglected to analyse their studies in different research areas and also expertise within research areas (see chapter 2.5.2.1).

Taken together, our meta-analysis is more limited than Karelaia and Hogarth's meta-analysis (2008), hence, only their results on one shot learning are comparable to our results (Table 2, p. 412). They showed that judgment achievement is moderate (r_a = .41) and the other values are high (G = 63; R_s = .80; R_e = .71) except one low value (C = 0.07). However, in the following if we refer to the work by Karelaia and Hogarth (2008), we focus on this aspect of their work. In addition, they focus also on expert models, as this is not the scope of our analysis we refer to chapter 7.

In conclusion, there are two meta-analyses done in the framework of the *SJT*, neither one used a psychometric Hunter-Schmidt approach or focused on the difference of judgment-achievement values in research areas. Furthermore, also the comprehensive data base used by the Karelaia and Hogarth (2008) meta-analysis left some questions open, such as what kind of cues or criterions influenced judgment achievement. Hence, a more detailed analysis is urgently needed.

2.5.1.3 Research on the Lens-Model components

A variety of research on the statistical and empirical behaviour of the components of the *LME* has been carried out over the last 40 years. For an overview see Cooksey (1996, p. 216). This research is designed to investigate the interrelationships among the components and their proper interpretation(s). For instance, a number of studies have found evidence that the reliability of judgment is lower for less predictable tasks (i.e. Brehmer, 1976; Harvey, 1995). Additionally, as Lee and Yates (1992) showed the lens-model statistics (i.e. *G*) decrease with an increasing number of cues.

In summary, a lot of designed studies have been conducted, but no research, apart from the two noted meta-analyses, has ever compared the already published studies and compared the actual components of the *LME*. Therefore, our research results can also be important from a more theoretical point of view to validate the existing research on the *LME* components.

2.5.2 Related to other judgment and decision-making approaches

There are already several reviews and meta-analyses on judgment achievement (for pros and cons of reviews and meta-analyses, see chapter 4.4).

An important starting-point for several reviews on judgment achievement is Meehl's work from 1954. He compared clinical-judgment makers' accuracy with the model of actual formulas (i.e. mechanical, statistical, actuarial), and as already introduced, this work clearly shows the advantages of actual formals compared to clinical judgments. Interestingly, the experience levels as well as the cues given to a judge did not affect the superiority of the actual method. However, Meehl (1954) showed that the model accomplished an almost better judgment than the experts. The recent review article on this topic by Grove et al. (2000) confirms Meehl's conclusion that mechanical prediction of human behaviour is equal or superior to clinical prediction methods. In addition, an overview focusing only on consulting area and judgment achievement supported the superiority of the model (see Aegisdottir et al., 2006). It must be mentioned, however, that reviews on this field only focus on aggregated data and ignore single judgment and decision-maker.

2.5.2.1 Expertise in research areas

It must be mentioned that in the review by Grove et al. (2000), there is also a trend toward greater advantage for expert models in the medical and forensic field as opposed to educational or finance settings, implying that variations in judgment achievement depend on subject area (i.e. research areas, domains). In addition, Armstrong's review (2001) implies differences in research areas in judgment achievement.

Furthermore, Shanteau (2002) supports the view of domain differences in judgment research and claims for research on expertise knowledge within different domains.

Altogether, although there are several meta-analyses and overviews on lens-model research, none of them take different research areas into account. One exception is the work by Dhami et al. (2004); as Dhami et al. (2004) also coded their studies according to research areas, but never published this aspect.

To summarize: the lens-model approach is a domain-independent approach, treading all areas equal. However, as introduced, research on judgment achievement focuses on domain-differences (e.g. Armstrong, 2001; Grove et al., 2000).

2.6 Summary of chapter 2

After the introduction of an array of different judgment and decision-making theories, their drawbacks led us to focus on the *SJT*. Because the *SJT* is defined as domain-independent, there are no overviews considering either different research areas or expertise knowledge within the research areas. Hence, we included these missing aspects in the *SJT* in our research questions, which are introduced in the following.

3 RESEARCH QUESTIONS

To reveal why some people are more accurate than others, numerous studies have applied the *LME* to judgments. Although the initial goal of the *LME* was to conduct comparative studies among judges and among situations, no complete meta-analysis according to Hunter and Schmidt (2004) of the *LME* has been conducted that gives an overview of judgment achievement between and across persons dependent on judgment tasks. To provide a comprehensive overview of judgment achievement in the framework of the *SJT*, we want to find out what the judgment achievement of persons actually is. Therefore, we will first focus on studies using an idiographic research approach. The following question interested us for our meta-analysis across persons:

1) What is the actual value of *judgment achievement* expressed as the correlation between judgments and criterion? For example, how good are a teacher's reading-achievement estimations actually?

The following questions are raised because the judgment achievement in the included studies is described by additional components of correlations of the *LME:*

2) What is the actual value of *error-free judgment achievement (knowledge)* expressed as the correlation between judgments and criterion, assuming the environment (i.e. judgment task) is perfectly predictable and the judges are perfectly consistent? For example, how good could a teacher's reading-achievement estimation actually be, assuming that the reading achievements of the students are perfectly predictable and the teacher's estimations are perfectly repeated?
3) What is the actual value *consistency* expressed as the correlation between cues and judgments? For example, how similar are repeated reading-achievement estimations of a teacher actually?

4) What is the actual value of *environmental predictability* expressed as the correlation between cues and criterion? For example, how well could information used by a teacher predict the actual reading achievement of students?

We began our analysis at the idiographic level, treating the data of the individuals as unique systems, before searching for commonalities in their judgment achievements.

An important issue that cuts across all the above questions is the influence of moderating factors, such as task characteristics on judgment achievement. Consequently, the meta-analysis will be repeated and checked as to whether the correlations of the components also are held for possible moderator as:

a) *Analysis unit,* such as individual (i.e. idiographic) or across individuals (i.e. nomothetic) analysis: Are there any differences, indicating an ecological fallacy? (see also chapter 2.3.4.1).

b) *Applied research area* (i.e. medical, business, educational and psychological science, or other research areas): How accurate are individuals' judgments in different domains in comparison to judgments across research areas? Are there any differences? What are the underlying reasons for accuracy or inaccuracy? (see also chapter 2.5.2.1).

c) If the *expertise* within one area may influence judgment achievement: For example, are medical experts' judgment achievements lower than teachers'? (see Shanteau, 2002).

4 METHODS

In the following section, the literature search and the resulting data base are described. Two control strategies and nine exclusion criteria leading to the final data base are also explained, followed by the coding procedure used. In addition, the Hunter and Schmidt (2004) method of meta-analysis is shortly introduced.

4.1 Literature search

4.1.1 Search strategies

The literature search is restricted to those articles published from 1964 until August 2008. In 1964, several publications (Hammond et al., 1964; Hursch et al., 1964; Tucker, 1964) led to the development of the *LME,* which provided the stimulus for many later studies. These studies including judgment achievement decomposed by the *LME* represent our data base. The relevant studies were identified by the five different search strategies (for an overview, see Figure 5) described in the following:

1) The mailing-list of the Brunswik Society was used to inform the society members about our research goal. Furthermore, the society members were asked whether they know any important literature in this research area.

2) The Brunswik Society Newsletter (1991-2007)[2] was searched for references. In addition, we used the annual Newsletter to inform about our project and to call for studies since 2006 (see Kaufmann, 2006, 2007; Kaufmann & Sjödahl, 2006; Kaufmann et al., 2007, 2008).

[2] http://www.brunswik.org/newsletters/index.html

3) To get an overview, the search was based on important articles and books in this research area. We used the following articles:

- Hammond, Hursch and Todd (1964)
- Hursch, Hammond and Hursch (1964)
- Tucker (1964)
- Castellan (1972)

These articles were used for our search in the *Web of Science* database. *Web of Science* allows the user to conduct "cited reference searching". Thus, the search started with the key article on *LME* (Tucker, 1964). Later articles referencing it will be identified shortly. The same strategy is also used with other important articles in the research area.

As previously noted, we also used *books* for our search, such as:

- Cooksey (1996)
- Hammond (2000)
- Hammond and Stewart (2001)

More precisely, we used the reference list of the books for our search. We also consulted google's book data (http://books.google.com) and searched for cited research.

4) For our search in 10 different data bases, seven different key-words were used. The key-words we got from our previously searched articles are suitable for our meta-analysis. To include both British and U.S. articles, we also considered the English expression, which is sometimes slightly different; "judgment", for instance in American English, and "judgement" in British English. As one data base is in German, the German equivalents of the

English expressions are used (see the parentheses in the following key-word list):

1. Social judg(e)ment theory (Soziale Urteilstheorie)
2. Lens model equation (Linsen Modell Gleichung)
3. Lens model analysis (Linsen Modell Analyse)
4. Lens model (Linsen-Model)
5. Judg(e)ment achievement (Urteilsleistung)
6. Idiographic approach (Idiographischer Ansatz)
7. Judg(e)ment accuracy (Urteilsgenauigkeit)

These key-words were used in the following data bases:

- *ERIC:* Educational Resources Information Center, since1966

- *EricOnline:* Educational Resources Information Center, online

- *PSYNDEX:* German literature on psychology, since 1977

- *PsychInfo:* Previously PsycLit., since 1806, international literature on psychology

- *Web of Science:* Journals on humanity, social sciences, and natural sciences

- *WISO-Net:* Literature on business science and social sciences

The *EricOnline* data base includes references to unpublished reports and conference papers in addition to published works.

Hence, using *EricOnline* leads to the prevention of publication bias (see chapter 4.6). The previously introduced key-works were also used in the following online search engines:

- *Google:*	*http://www.google.com*
- *Google Scholar:*	*http://scholar.google.com*
- *Yahoo:*	*http://www.yahoo.com*
- *Social Science Research Network:*	*http://www.ssrn.com/*

The resulting literature was first scrutinized for key-words in the title and in the abstract. Then, we checked, whether the reference list cited Tucker, Hammond or Brunswik. For a comprehensive overview including data and results of our literature search and further information about the used data base we refer to the Appendix B: Tables 1, 2, 3.

5) Finally, we also created different google-alerts with the key-words in order to be informed about the ongoing work on this subject, such as newly initiated projects or publications.

The literature review is therefore as up-to-date as possible. However, new articles on this subject have undoubtedly appeared since the last up-date of our literature search in August 2008.

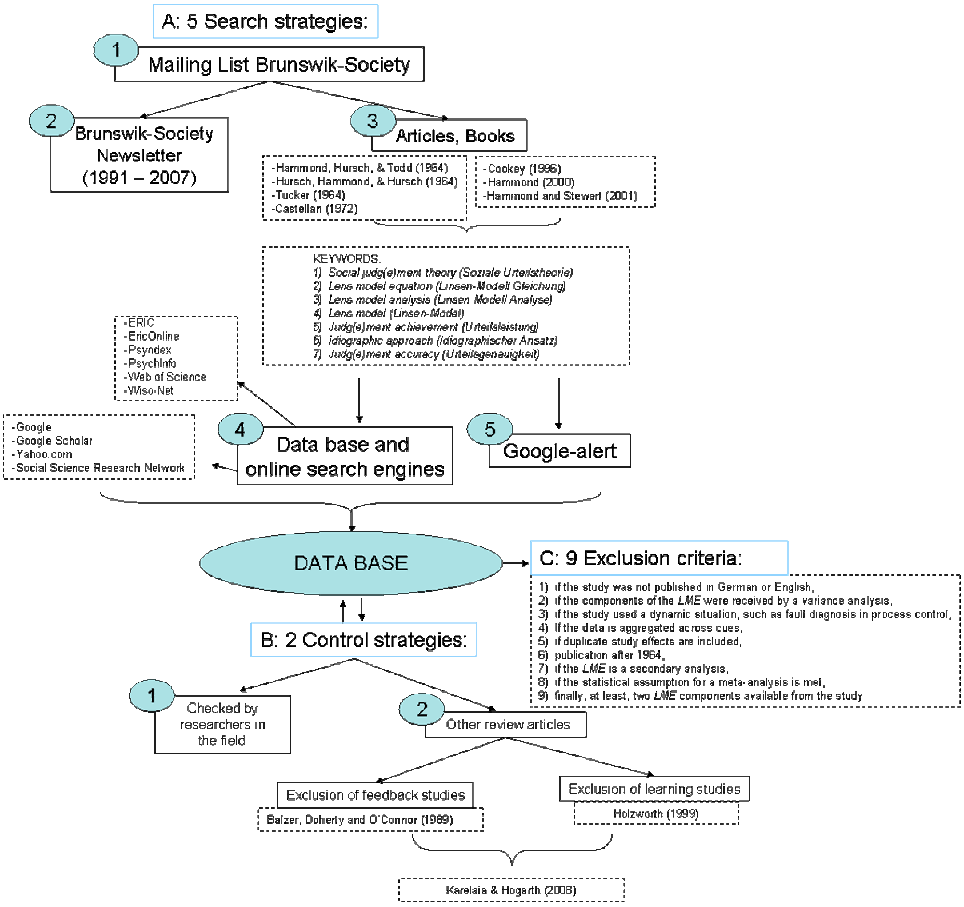

A: 5 Search strategies:
1
Mailing List Brunswik-Society
2
Brunswik-Society Newsletter (1991 – 2007)
3
Articles, Books
-Hammond, Hursch, & Todd (1964)
-Hursch, Hammond, & Hursch (1964)
-Tucker (1964)
-Castellan (1972)
-Cookey (1996)
-Hammond (2000)
-Hammond and Stewart (2001)
KEYWORDS.
1) Social judg(e)ment theory (Soziale Urteilstheorie)
2) Lens model equation (Linsen-Modell Gleichung)
3) Lens model analysis (Linsen Modell Analyse)
4) Lens model (Linsen-Model)
5) Judg(e)ment achievement (Urteilsleistung)
6) Idiographic approach (Idiographischer Ansatz)
7) Judg(e)ment accuracy (Urteilsgenauigkeit)
-ERIC
-EricOnline
-Psyndex
-PsychInfo
-Web of Science
-Wiso-Net
-Google
-Google Scholar
-Yahoo.com
-Social Science Research Network
4
Data base and online search engines
5
Google-alert
DATA BASE
C: 9 Exclusion criteria:
1) if the study was not published in German or English,
2) if the components of the LME were received by a variance analysis,
3) if the study used a dynamic situation, such as fault diagnosis in process control,
4) If the data is aggregated across cues,
5) if duplicate study effects are included,
6) publication after 1964,
7) if the LME is a secondary analysis,
8) if the statistical assumption for a meta-analysis is met,
9) finally, at least, two LME components available from the study
B: 2 Control strategies:
1
Checked by researchers in the field
2
Other review articles
Exclusion of feedback studies
Balzer, Doherty and O'Connor (1989)
Exclusion of learning studies
Holzworth (1999)
Karelaia & Hogarth (2008)

Figure 5. A flowchart of the literature-search model including the five search strategies, two control strategies, and the nine exclusion criteria.

4.1.2 Results of our literature search

In line with Dhami et al. (2004, p. 964) our literature search showed that the lens model was used to study judgment achievement or judgment accuracy, multiple-cue probability learning, or cognitive feedback.

Most studies resulting from our search applied the Lens Model to investigate multiple-cue probability learning (see Holzworth, 1999).

In addition, many studies applied the Lens Model to study the effectiveness of cognitive feedback. Feedback, for example cognitive feedback, consists of three types, which possibly depend upon the judgment design implemented: *Task information* (summarizing how cues relate to some distal criterion), *cognitive information* (summarizing how cues relate to the person's judgments), and *functional validity* information (summarizing the links between ecology and judgment measurement and models, see Cooksey, 1996, p. 367). Consequently, each type of cognitive feedback also represents three types of different studies. For an overview see Balzer et al. (1989), who also conclude that "because there are differences among cognitive feedback studies, any attempt to generalize across these studies must attempt to take these differences into account" (p. 414).

Finally, our literature search showed that only a minority of studies applied the Lens Model to investigate judgment achievement. This result of our literature search is in line with Dhami et al. (2004), as they conclude:

> "…the relative neglect of the study of achievement by neo-Brunswikians is surprising in light of Brunswik's (1943, 1952) emphasis on achievement as the topic of psychological research, …". (p. 968)

Consequently, feedback and learning studies were excluded, since the focus of this meta-analysis is judgment accuracy across different situations and contexts. For a general meta-analysis, in which multiple-cue probability learning and feedback studies were also considered, see Karelaia and Hogarth (2008).

4.1.3 Control strategies

In order to exclude all feedback and learning studies and include only judgment achievement studies, our data base was checked by two major strategies (for an overview of our control strategies, see Figure 5):

1) Checked by researcher in the field:
   The included studies were mailed to relevant researchers in the field (Stewart, Tickle-Degnen) for scrutiny. For all studies published since 2000, we also actively tried to contact the first author. Hence, we wrote to *Cooksey* (19.05.07) and *Trailer* (24.05.08) and asked for the results of their current work. Furthermore, three studies were sent to *Doherty,* who is a co-author of them (23.06.08, see Gorman et al., 1978; Roose & Doherty, 1976; Steinmann & Doherty, 1972). *Werner* and *Bernieri* were also contacted (02.07.08). Thanks to Werner we found an unpublished dissertation by Lehman (1992). We also used these contacts to scrutinize the coding of the studies (see chapter 4.2).

2) In addition, the database was checked by other review articles to determine
   a. whether *all feedback studies* were excluded by the review by Balzer et al. (1989) and

   b. whether *all learning studies* were excluded by the annotated bibliography by Holzworth (1999).

   c. Naturally, we also took advantage of the fact that Karelaia and Hogarth (2008) conducted a meta-analysis as well. Therefore, our resulting literature base was checked after the publication of the article by Karelaia and Hogarth to see whether we actually excluded *all learning and feedback studies*.

After these control checks, we found no need to include or exclude any feedback or learning studies in our meta-analysis.

4.1.4 Excluded achievement studies

To compare the included studies focused on judgment achievement, nine exclusion criteria were applied to these studies (see Figure 5). These nine exclusion criteria are explained in detail in the following.

First, our data base was restricted to studies in German and English. We must mention, however, that, our data base included only English publications in the end. We found only a minor sample of German studies and had to exclude them, as they did not meet all inclusion criteria (e.g. Wittmann, 1985, see below).

Secondly, we only used studies containing a regression analysis in order to exclude any heterogeneity between the studies resulting from different analysis methods.

Thirdly, we also excluded studies in dynamic situations (Kirlik, 2006). Hence, only studies with stable situation representations are included, so as to prevent that the differences in judgment achievement will result from the representation of the situation.

Fourthly, as mentioned above, we excluded studies aggregated across cues instead of individuals, in order to prevent any aggregation bias (e.g. Wittmann, 1985).

Fifthly, we checked for data included twice in the data base, to prevent double counting – which was not the case (see Wood, 2008).

Sixthly, we also excluded the study by Hammond (1955). Although he was the first to apply the Lens Model to judgments, he could not include the *LME,* because it hadn't been developed yet. Even after the publication by Tucker (1964), a number of studies applied the Lens Model to judgments without using the *LME* (see Lyons, Tickle-Degnen, Henry, & Cohn, 2004; Tickle-Degnen & Lyons, 2004). For a comprehensive overview of the excluded studies, see Appendix D: Table 1.

Seventhly, we only included studies, in which the *LME* analysis was used as a primary analysis. For instance, the study by Camerer (1981), who calculated the *LME* components of already published studies and meta-analysed these studies. Hence, we see this study as a secondary analysis. In addition, Camerer's focus was bootstrapping modelling (in the following we use the term expert model), which is not the focus of this work. These arguments lead us to the exclusion of the analysis by Camerer (1981) in a first step, but we checked and compared our results with Camerer's study in our robustness analysis (see Kaufmann & Athanasou, 2009).

Eighthly, 35 studies meet our inclusion criteria without regarding statistical *presumption for a meta-analysis*. Hence, studies were excluded because of a lack of data for conducting a meta-analysis (e.g. Goldberg, 1970; Tape, Heckerling, Ornato, & Wigton, 1991).

Finally, we would like to mention that two components, namely achievement and task predictability, should be available from the studies.

Consequently, a total of 31 studies met the inclusion criteria (see Tables 5 and 6).

4.2 Coding studies

First, all essential information according to the research questions and potential moderating variables as well as information by other reviews articles in the fields (e.g. Armstrong, 2001; Cooksey, 1996; Dhami et al., 2004) were included to a first version of a coding scheme, which was then adopted by coding the first articles.

The following major coding categories were included:

a) Study ID: A unique number for every study
b) Study characteristics (e.g. publication year, author)
c) Characteristics of the research participants (e.g. students)
d) Characteristics of the judgment tasks (e.g. number of cues)
e) Effect size: The effect size was collected or calculated for each task. Consequently, where it was possible, we broke

down multivariate judgment tasks to univariate ones, in order to achieve a more precise analysis. In addition, we calculated missing *LME* components (for the formulas, see Appendix C).

Consequently, all essential information was extracted from the selected studies and included in a SYSTAT (2000) file (for a detailed coding scheme, see CD, coding.doc).

4.2.1 Coding reliability

Finally, the studies were coded by the author alone, therefore, it was not possible to calculate the interrater reliability, such as cronbach's alpha.

However, the coding of the studies was checked by two control strategies: a) by the authors of 10 studies, and b) by other review articles (e.g. Karelaia & Hogarth, 2008). These control strategies are described in detail in the following.

Our study coding was first checked by the authors. Thanks to *Athanasou* (Athanasou & Cooksey, 1996), *Werner* (Cooper & Werner, 1990; Lehman, 1992; Werner et al., 1989; Werner et al., 1983), *Doherty* (Gorman et al., 1978; Roose & Doherty, 1976; Steinmann & Doherty, 1972) and *Stewart* (Stewart, 1990; Stewart et al., 1997) 10 studies were controlled by the authors themselves.

We then reported the agreement between our database and other reviews (see Table 2). To simplify matters, Table 2 is separated into two sections. Hence, before representing the agreement of *LME* values, we reported the agreement of the different study characteristics.

Table 2

Agreement of our data base with study characteristics and LME values by other reviews

Review	Number of overlapping studies	Comparison characteristics: Study characteristics/ *LME* components	Agreement
Armstrong (2001)	4 (7, 8, 9, 16)	Number of cues, number of judgments	100%
Ashton (2000)	2 (2, 5a)	Number of judgments	100%
Karelaia & Hogarth (2008)	19 (2, 3, 4, 7, 8, 11, 12, 13, 15, 16, 17, 18, 19, 20, 21, 22, 27, 28, 29)	Number of judges, number of judgments, number of cues, experience level	92%
		r_a, G, R_s ,$R_{e,}$ C	88%
Aegisdottir et al. (2006)	2 (16, 18)	r_a, R_e	100%
Grove et al. (2000)	3 (9, 16, 18)	r_a, R_e	50%

As you can see in Table 2, we first compared our data with the review by Armstrong (2001). The first row shows the review, the second row contains the total number of overlapping studies and the study numbers in parentheses. These numbers are the same you will find in Tables 5 and 6. To give an example, number 17 represents the study by Athanasou and Cooksey (2001). However, four studies (Ashton, 1982; Goldberg, 1976; Roose & Doherty, 1976; Wiggins & Kohen, 1971) are checked for two study characteristics: number of judgments and number of cues (see third row). We had a 100% agreement. In addition, our data base was compared to the article by Ashton (2000) for the study characteristic number of judgments. The data of two overlapping studies (Einhorn, 1974; Levi, 1989) is completely identical. Finally, as mentioned before, we utilised the fact that Karelaia and Hogarth (2008) also conducted a meta-analysis: We selected all judgment achievement studies from their data base and compared them to ours. There were 19 studies used in both meta-analyses, which could be distinctly identified, as the same studies. Therefore, it is a conservative estimation, as several studies are not clearly identifiable, as for instance the publication year is different (see Stewart, 1989, 1990). First of all, we compared as to whether our classification of achievement studies was comparable, and there was only one misclassification. Hence, we achieved a 95% agreement. Secondly, we checked the separation of the studies or the number of judgment tasks included by single studies. Instead of 30 judgment tasks, we used only 28. Although there are some differences, if a final meta-analysis is conducted over this subsample of studies, there is almost no difference between this subsample of databases. However, there are differences if we separate some tasks (Gorman et al., 1978; Stewart, 1997), as Karelaia and Hogarth's did not and reverse (Cooper & Werner, 1990; Harvey & Harries, 2004; Wright, 1979). But, most of the studies are coded in the same way (74%). Thirdly, we compared the four study characteristics number of judges, judgments, cues, and expertise level of the judges (see Appendix D: Table 3). The data base agrees with 92%. In addition, we checked the *LME* values, and our high agreement was confirmed (88% for details, see Appendix D: Tables 4, 5). Furthermore, we checked by Aegisdottir et al. (2006), and the *LME* components r_a and R_e resulted in a 100% agreement. Grove et al. (2000) were also used for checking our *LME* components r_a and R_e. However, our component R_e, was not comparable with the mechanical accuracy of Grove et al. (2000). Hence, we reached an agreement of merely 50%.

To summarize: Our data base was compared with five reviews, and except for the review by Grove et al. (2000), a high agreement was found.

4.3 Description of the studies

The resulting 31 studies of the literature search are described in Tables 5 and 6. The tables display the name(s) of the author(s) and the publication year along with the study characteristics that were used to describe the studies and the study results. Furthermore, the studies are separated into five different research areas. Within each research area, they are ordered according to the experience of the judges. And, finally, within an experience level, the studies are ordered according to the number of used cues.

However, in the following, we will describe the complete data base for our meta-analysis. Firstly, we will compare them to important review articles in the field. Then, we will focus on the database in relation to the type of publication. Thirdly, we will introduce the main study characteristics, such as research approach and research area. Furthermore, we will mention further study characteristics like the number of cues and the type of judgment criterion.

4.3.1 In relation to other reviews on judgment achievement

The following Table 3 gives an overview of the number of overlapping studies from our meta-analysis with other reviews in the field. However, only the nomothetic data base was considered, as we weren't aware of any idiographic-based review on judgment achievement.

In the first row, the table shows the study, then the number of overlapping studies. As already mentioned, the number is the same as in Tables 5 and 6. Finally, in the last row, the total number of the overlapping studies can be found.

Within the *SJT* approach there is a great overlap between our data base and the study by Karelaia and Hogarth (2008). However, the differences were already mentioned. In contrast, there is only a small overlap with the study by Dhami et al. (2004), whereas this review wasn't a meta-analysis. Beyond the *SJT* approach, there seems only a small overlap between our studies and other review articles from two to four studies, although two of these are also meta-analysis. However, if the total number of studies is considered, it becomes clear that the greatest overlap is with the Armstrong's review (2001).

To summarize: Although there are differences in the coding of the data (see chapter 4.2.1), the greatest overlap of our data base within the *SJT* is with the meta-analysis by Karelaia and Hogarth (2008), and beyond the *SJT* approach with the Armstrong review (2001).

Table 3

Review articles including studies overlapping with our meta-analysis, ordered by the total number of overlapping studies. Firstly, within the SJT, and then beyond the SJT approach.

Reviews	The number of studies overlapping with our meta-analysis	Total number of overlapping studies
Within the SJT approach		
Karelaia & Hogarth (2008)	2, 3, 4, 7, 8, 11, 12, 13, 15, 16, 17, 18, 19, 20, 21, 22, 27, 28, 29,	19
Dhami et al. (2004)[a]	3, 15, 19, 24, 29	5
Beyond the SJT approach		
Armstrong (2001)	7, 8, 9, 16	4
Grove et al. (2000)	9, 16, 18	3
Aegisdottir et al. (2006)	16, 18	2

Note. [a]see reference list at the Brunswik Society homepage.

4.3.2 Journal of the publications

All articles included in our meta-analysis are written in English and were published within the last 40 years. It must be mentioned that the study by Lehman (1992) is an unpublished doctoral thesis. However, to report the complete data base, we included this study with the publication year or finishing of the thesis 1992 (see Lehman – according to our personal communication, the thesis will be published in the coming year).

The oldest study was published in the *Journal of Personality and Social Psychology* in 1971. The two most recent articles were published in 2004 (see Table 5, 6). As you can see in the following Figure 6, eight of the 31 articles were published in the 1970s, 11 were published in the 1980s, seven were published in the 1990s, and five between 2000 and 2005. All articles apart from two (Athanasou & Cooksey, 2001; MacGregor & Slovic, 1986) were published in ISI-indexed journals in May 2006. Journals that have published more than one study are:

1) *Organizational Behavior and Human Decision Processes*[3]
   (see Einhorn, 1974; Goldberg, 1976; Gorman et al., 1978; Roose & Doherty, 1976; Steinmann & Doherty, 1972; Stewart et al., 1997; Wright, 1979),
2) *Journal of Personality and Social Psychology*
   (see Bernieri et al., 1996; Wiggens & Kohen, 1971)

Thus, articles in this research area were mainly published in the 1980s in the *Journal Organizational Behavior and Human Decision Processes*.

Finally, Figure 6 clearly gives evidence of the continuing value of the *LME* for judgment analysis since more than 40 years.

[3] Organizational Behavior and Human Performance before 1985

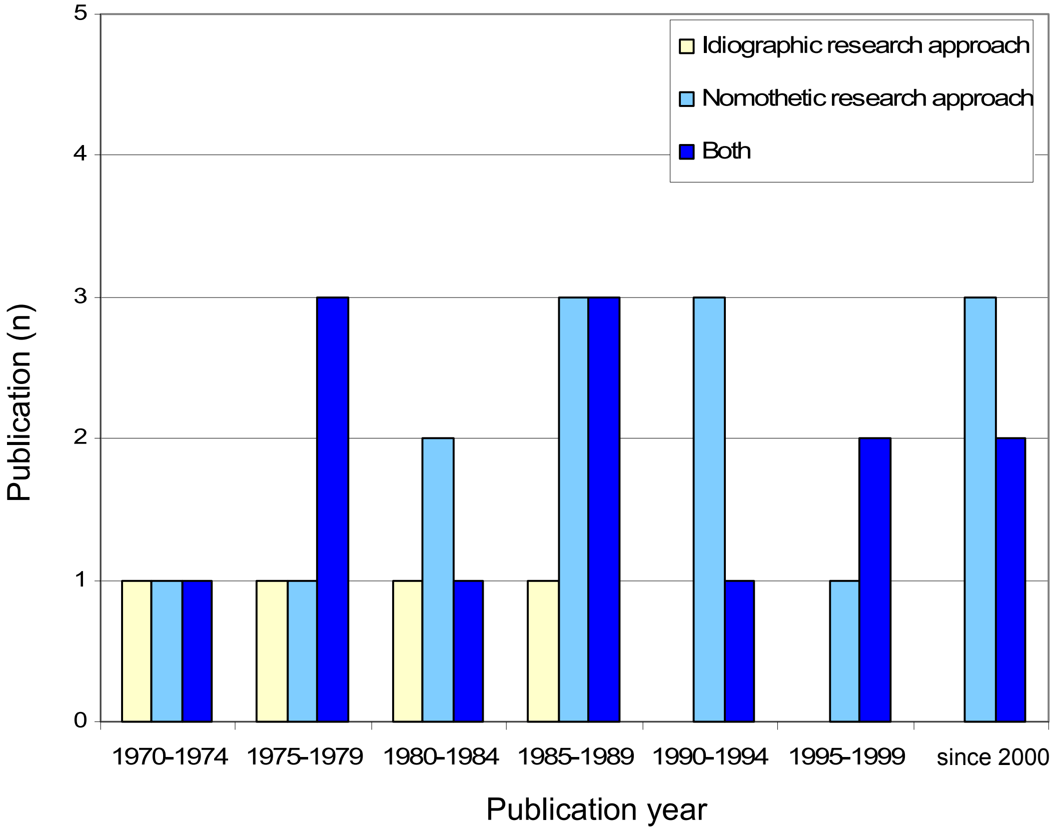


Figure 6. The number of publications separated by their research approach – idiographic versus nomothetic, or both.

4.3.3 Research approaches

As you can also see in Figure 6, the study characteristic applied research approach is presented. Studies are considered idiographic, if they generate a correlation between judgments and criterion within each individual before obtaining any aggregate or nomothetic measures of relationship.

Furthermore, the study characteristic used research approach is presented in the last column in Tables 5 and 6. If an asterisk is found in this last column, the *LME* is applied only to individuals (i.e. idiographic approach), and if a + is added, also the nomothetic approach is considered.

However, as you can see in Figure 6, most studies used a nomothetic research approach and seldom an idiographic approach. In addition, most studies are not considered or compared both research approaches. Although the data base is enlarged because we added some studies (e.g. Lehman, 1992), in our 2007 data base (see Kaufmann, 2007) the conclusion is that the idiographic research approach is clearly neglected in recent *LME* research.

4.3.4 Research areas

Tables 5 and 6 present the characteristics of the studies resulting from our comprehensive literature search. A total of 31 studies included 49 different judgment tasks. Different tasks within a study are symbolized by a roman numeral. In these 49 judgment tasks, 1151 judgment achievements were made by 1055 subjects. Hence, 96 judgment achievements of 68 persons were analysed for more than one task. As you can see in Table 4, two studies used the same judges for analysing four tasks, one study for three tasks, and, finally, three studies used their judges for the analysis of two tasks. It must be mentioned, however, that the studies reported in Table 4 include experts as well as non-experts and come from different research areas as introduced in the following.

Table 4

The number of judges in studies analysed judges more than once ordered by the amount of judgment task for each judge

Study	Number of analysed judgment tasks for each individual (total number of judges)
Gorman et al. (1978)	4(8)
Stewart (1997)	4(4)
Nystedt & Magnusson (1975)	3(4)
Einhorn (1974)	2(29)
Kim et al. (1997)	2(3)
Cooksey et al. (1986)	2(20)

In addition, in Tables 5 and 6 the judgment tasks are separated into five research areas. Within the research areas, tasks are arranged according to the amount of cues considered. Furthermore, in their categories the studies are also sorted by the experience level as experts or students. In the following, the study characteristics within each research area are described in more detail:

1) Six studies applied to the *medical science* category include a total of 221 clinical oriented educations (e.g. clinical psychologists) and 258 analyzed judgment achievements for 10 judgment tasks. The studies applied to medical science include those with the overall lowest and the overall highest number of judgments and represent the only category including only experts. Furthermore, this category contains the most studies with an idiographic approach, including 95 analyzed judgment achievements of 58 individuals in eight different tasks. We would like to mention the small sample in the first study by Einhorn (1974), which includes only three pathologists. However, these three pathologists are the only ones in these categories who made their judgments on real biopsy slides, representing a more natural situation than the commonly used patient profiles. In addition, in our opinion, this is also the only prognostic task compared to the remaining diagnostic tasks.

2) Table 5 contains eight studies applied to *business science*. As three person's judges' two tasks, 40 analyzed judgment achievements by 37 persons in five different tasks reported individual data. The study by Wright (1979) analyzed only the five most accurate judgments of the 47 included persons at the idiographic level. However, if also studies with a nomothetic research approach are considered, eight studies report nine judgment tasks judged by 236 persons. In addition, experts in the included studies in business science are managers, bank loan

officers, and security analysts. Studies applied to business science have the largest range of the number of cues (seven to 64). In two studies the number of cues is unknown. Furthermore, all judgments were based on paper profiles.

3) In Table 5, three studies applied to *educational science* are included. 136 persons judged four different judgment tasks in this category. Most of them (72%) are students from the study by Wiggins and Kohen (1971). Once again, all judgments were based on paper profiles. The study by Cooksey et al. (1986) includes a multivariate lens model design (see chapter 2.4.1.2). However, they analysed judgment by single criteria, and therefore, two tasks are available. In addition, in this category, 58 judgment achievements by 38 individuals are available.

4) Table 5 includes eight studies of 225 persons, which are applied to *psychological science*. 59 persons are experts, as they all have experience in their judgment tasks, in contrast to the 166 students included in psychological science studies. In this category, 43 judgment achievements of 19 individuals (11 experts) made 1580 judgments in six different tasks are available.

5) Because six studies could not be categorized accurately, the *other research area* category was created. Therefore, Table 6 contains an additional column for the applied research area. In this category, studies like weather forecasts or judgments of time taken to run a marathon are included. This category includes the individual data base on nine experts or meteorologists in five different tasks. Individual data is also available from 97 of the 228 students who judged 11 tasks. The study by Stewart et al. (1997) is the only one which compares non-experts and experts across

four meteorological tasks. Furthermore, this study is also the only to analyse the judgment achievement retrospectively.

In summary, 1055 persons judged 49 tasks across 31 studies: 21% of persons were included in studies applied to medical science, 22% in studies applied to business science, 13% in studies applied to educational science, 21% to psychological science, and 23% were applied to other research areas. As can be seen in Tables 5 and 6, the first application of the *LME* was in disciplines such as medical or business science. By and by, the application of the *LME* was expanded to other research areas, such as psychology or meteorology. In addition, the great majority of studies used students, often enrolled in psychology classes. An exception in our review is the category of medical science. In medical science, all studies include experts.

With a look across judgment tasks you can see that 80% (eight judgment tasks, see Figure 7) in medical science used an idiographic research approach. However, the application of the idiographic research approach decreases to a level of 50% in psychological studies.

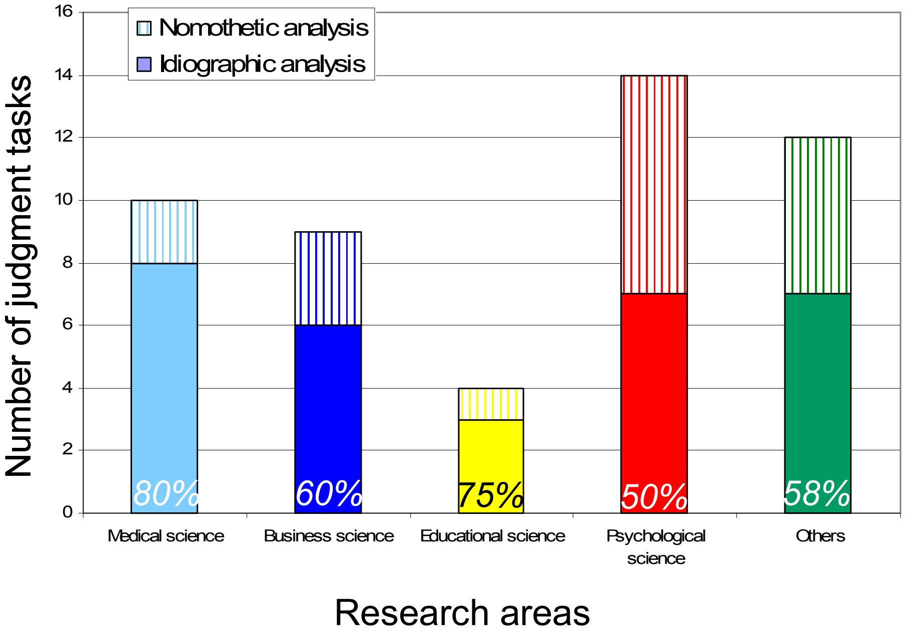


Figure 7. The precental distribution of idiographically vs. nomothetically analyzed judgment tasks within the five research areas.

In the following, several task characteristics across research areas are illustrated.

4.3.5 The number of cues

As you can see in Tables 5 and 6, the number of cues is also consider in each study, although in two studies in business science the number of cues is unknown.

Although the Fast and Frugal Heuristic Approach (Gigerenzer et al., 1999) assumes that decision makers use only few cues, only five studies used less than five cues. One of these five studies applies to in business (Wright, 1979), medical (Nystedt & Magnusson, 1975), and psychology (Szucko & Kleinmuntz, 1981) science. In addition, two of these five studies are in the miscellaneous research area (MacGregor & Slovic, 1986; Steinmann & Doherty, 1972). Although four of the five studies applied an idiographic research approach, only one also uses a nomothetic research approach (Nystedt & Magnusson, 1975). Furthermore, in three of the five studies, the type of correlation is unknown (see chapter 4.3.7). The remaining 26 studies used more than five cues.

4.3.6 The criterion

Hammond (2007, p. 53) separated judgment achievement competence into accuracy in judging objects and events in the *physical environment* and accuracy in judging objects and events in the *social environment*. According to Hammond, judgment in our physical environment should be more accurate. One of his given reasons is that in the physical environment we receive clear and fast outcome feedback. On the other hand, in our social environment we could mistakenly judge our best friend to be honest, while it may take years to learn that he is not. However, all studies in Table 5 in medical science, psychological, and educational science are embedded in a social environment. On the other hand, studies in business science and the miscellaneous research area are all embedded in a physical environment. Consequently, if there is a tendency in our analysis that judgment achievements in medical, psychological, and educational science are more accurate than judgment achievements in business and the miscellaneous research area, Hammond's hypothesis that judgments in physical environments are more accurate would be supported.

However, we should also add the type of criterion. On the one hand, the criterion is subjective – for example the single physician's judgment in the study by LaDuca et al. (1988, see Table 5, to provide an overview, studies with a subjective criterion are marked with a triangle in the criterion column). On the other hand, the criterion is objective, as, for example actual temperature, which is measured by an instrument (see Stewart, 1997, Table 6).

4.3.7 The type of correlation

Usually, the correlation coefficient (e.g. Spearman's roh) was calculated, and a median or average coefficient for the sample was quoted. Studies with an unknown type of correlation are symbolized with r_0 (see last column in Tables 5 and 6). Furthermore, Spearman's roh is used in business, in psychological science or in other research area studies.

4.3.8 Summary of chapter 4.3

Our data base is heterogeneous and relates to different study characteristics, such as the number of used cues or research area. However, we also conclude that the idiographic approach was mostly neglected in recent studies (Kaufmann, 2007). Hence, the complementation of the nomothetic with an idiographic analysis is recommended in order to achieve a comprehensive overview on judgment achievement in the framework of the *SJT*.

Table 5

Studies included in our meta-analysis. The studies' characteristics are ordered according to the applied research area. Experts and non-experts are separated within the research areas, ordered also according to the number of cues included in one study

	Study	Number of judges	Number of judgments	Number of cues	Judgment task	Criterion	Results
a)	*Medical science, all included studies in our meta-analysis considered experts as subjects:*						
1)	Nystedt & Magnusson, 1975	4 clinical psychologists	38 patient protocols	3	Evaluate patients on three traits: *I:* Judgment on intelligence *II:* Judgment on ability to establish contact *III:* Judgment on control of affect and impulses	Rating on three psychologist tests	*I:* r_0 = .63 *II:* r_0 = .66 *II:* r_0 =.47 (*, +)
2)	Levi, 1989	9 nuclear medicine physicians	280 patient cases, 60 replications	5	Assess probability of significant coronary artery disease	Coronary angiography	r_s = .47 (*)
3)	LaDuca, Engel, & Chovan, 1988	13 physicians	30 patient profiles	5	Degree of severity (Congestive heart failure)	A single physician's judgment (▲)	r_0 = .62 (*)
4)	Smith, Gilhooly, & Walker, 2003	40 general practitioners	20 case profiles	8	Prescription of an antidepressant	Guideline expert (▲)	r_0 = .53
5a)	Einhorn, 1974 *Second study*	3 pathologists	*III:* 193 biopsy slides	9	Evaluate the severity of Hodgkin's disease	Actual number of months of survival	*III:* r_0 = -.001
5b)	Einhorn[4], 1974 *First study*	29 clinicians	*I:* 77 MMPI profiles *II:* 181 MMPI profiles	11	Judging of the degree of neuroticism-psychoticism	Actual diagnosis	*I:* r_0 = .16 *II:* r_0 = .19 (*, +)
6)	Speroff, Connors, & Dawson, 1989	123 physicians: 105 house staff, 15 fellows, 3 attending physicians	440 intensive care unit patients	32	Patients' hemodynamic status (Physicians' estimation)	The patient's actual hemodynamic status	r_c = .42

Note. ▲= subjective criterion. r_0 = type of correlation is unknown. (*) = idiographic approach (cumulating across individuals). (*, +) = both research approaches are considered.

[4] This publication contains two studies.

Table 5 *(continued)*.

	Study	Number of judges	Number of judgments	Number of cues	Judgment task	Criterion	Results
b)	*Business science, experts:*						
7)	Ashton, 1982	13 executives, managers, sales personnel	42 cases in a booklet	5	Predictions of advertising sales for *Time* magazine	Actual advertising pages sold	r_s = .75 (*, +)
8)	Roose & Doherty, 1976	16 agency managers	200 / 160 profiles	64 / 5	Predictability of success of life insurance salesman	One-year criterion for success	r_0 = .13 (*, +)
9)	Goldberg, 1976	43 bank loan officers	60 large industrial corporations profiles	5	Bankruptcy experience	Actual bankruptcy experience	r = .51
10)	Kim, Chung, & Paradice, 1997	3 experienced loan officers	119 financial profiles: *I:* 60 big firms, *II:* 59 small business firms	7	To judge whether a firm would be able to repay the loan requested	Actual financial data	*I:* r_0 = .53 *II:* r_0 = .58 (*, +)
11)	Mear & Firth, 1987	38 professional security analysts	30 financial profiles	10	Predicted security returns	Actual security returns	r = .12
	Students:						
12)	Wright, 1979	47 students	50 securities profiles	4	Price changes for stocks (from 1970 until 1971)	Actual financial data	r = .22 (*, +)
13)	Harvey & Harries, 2004 (1. experiment)	24 psychology students	40 profiles	Not known	Forecast sales outcomes	Actual sales outcome	r_0= .98
14)	Singh, 1990	52 business students	35 profiles	Not known	Estimates of the stock of a company	Actual realized values	r_0 = .84

Note. ▲= subjective criterion. r_0 = type of correlation is unknown. (*) = idiographic approach (cumulating across individuals). (*, +) = both research approaches are considered.

Table 5 *(continued).*

	Study	Number of judges	Number of judgments	Number of cues	Judgment task	Criterion	Results
c)	*Educational science, experts:*						
15)	Cooksey, Freebody, & Davidson, 1986	20 teachers	118 profiles of kindergarten children	5	*I:* Reading comprehension *II:* Word knowledge	*I-II:* Actual end-of-year scores of each student on the two tests (▲)	*I:* r_c = .56 *II:* r_c = .57 (*, +)
	Students:						
16)	Wiggins & Kohen, 1971	98 psychology graduate students	110 profiles	10	Forecast first-year-graduate grade point averages	Actual first-year-graduate grade point averages	r_0 = .33
17)	Athanasou & Cooksey, 2001	18 technical and further education students	120 student profiles	20	Deciding that students are interested in learning	Actual level of students' interest	r_0 = .31 (*, +)
d)	*Psychological science, experts:*						
18)	Szucko & Kleinmuntz, 1981	6 experienced polygraph interpreters	30 polygraph protocols	3-4	Truthful / untruthful responses	Actual theft	r_{pb}= .23 (*, +)
19)	Cooper & Werner, 1990	18: 9 psychologists 9 case managers	33 inmates' data forms	17	Forecast violence during the first 6 months of incarceration.	Actual violent behaviour within 6 months of imprisonment	r = -.01
20)	Werner, Rose, Murdach, & Yesavage, 1989	5 social workers	40 Admission data for psychiatric inpatients	19	Assess imminent violence in the first 7 days following admission	Actual outcome: violent acts in the first 7 days following admission	r = .18 (*, +)
21)	Werner, Rose, & Yesavage, 1983	30: 15 psychologists 15 psychiatrists	Case material for 40 male patients	19	Predicting patients' violence during the first 7 days following admission	Actual violence during the first 7 days following admission	r_s= .12

Note. ▲= subjective criterion. r_0 = type of correlation is unknown. (*) = idiographic approach (cumulating across individuals). (*, +) = both research approaches are considered.

Table 5 *(continued).*

	Study	Number of judges	Number of judgments	Number of cues	Judgment task	Criterion	Results
	Psychological science, students:						
22)	Gorman, Clover, & Doherty, 1978	8 students	75: *I, III:* 50 interviews *II, IV:* 25 paper-people	*II, IV:* 6 *I, III:* 12	Prediction of each student's scores in an attitude scale (*I, II*) and a psychology examination (*III, IV*)	Actual data: *I, II:* Attitude scale *III, IV:* Examination scale (▲)	*I:* r_0 = .23 *II:* r_0 = .05 *III:* r_0 = .46 *IV:* r_0 = .45 (*)
23)	Reynolds, & Gifford, 2001	*I:* 7 students *II:* 10 students *III:* 28 students	videotapes	7 8 9	To assess the intelligence I: Audio condition II: Visual condition III: Audio plus visual condition	Wonderlic Personnel Test (a brief intelligence test)	*I:* r = .22 *II:* r = .38 *III:* r = .30
24)	Bernieri, Gillis, Davis, & Grahe, 1996	*I:* 45 students *II:* 54 students	50 videotaped debates	*I:* 17 *II:* 24	Rapport judgments	Interactants self-reports context: *I:* Adversarial, or *II:* Cooperative (▲)	*I:* r = .19 *II:* r = .28
25)	Lehman, 1992	14 students	Case material for 40 male patients	19	Assess imminent violence in the first 7 days following admission	Actual outcome (violent acts in the first 7 days following admission)	r = .24 (*, +)

Note. ▲= subjective criterion. r_0 = type of correlation is unknown. (*) = idiographic approach (cumulating across individuals). (*, +) = both research approaches are considered.

Table 6

Miscellaneous studies included in our meta-analysis. The studies' characteristics are ordered by expertise and also according to the number of cues included in one study

	Study	Number of judges	Number of judgments	Number of cues	Judgment task	Criterion	Research area	Results
e)	*Miscellaneous research area, experts:*							
26)	Stewart, 1990	7 meteorologists	75 radar volume scans (25)	6	Assess probability of hail or severe hail	Observed event	Meteorology	r_0 = .43 (*)
	Both, experts and students:							
27)	Stewart, Roebber, & Bosart, 1997	4: 2 students 2 experts	*I:* 169 forecast days *II:* 178 forecast days *III:* 149 forecast days *IV:* 150 forecast days	12 13 24 24	24-h maximum temperature forecasts 12-h minimum temperature forecasts 12-h precipitation forecasts 24-h precipitation forecasts	*I, II:* Actual temperature *III, IV:* Actual precipitation	Meteorology	*I:* r_0 = .96 *II:* r_0 = .96 *III:* r_0 = .74 *IV:* r_0 = .71 (*, +)
	Students:							
28)	Steinmann & Doherty, 1972	22 students	192: 2 sessions with 96 judgments	2	To decide from which of two randomly chosen bags a sequence of chips had been drawn	A hypothetical „judge" (▲)	Other	r_o = .65 (*)
29)	MacGregor & Slovic, 1986	*I:* 25 students *II:* 25 students *III:* 26 students *VI:* 27 students	*I - IV:* 40 runner profiles	4	Estimation of the time to complete a marathon	Actual time to complete the marathon	Sport	*I:* r = .42 *II:* r = .63 *III:* r = .39 *VI:* r = .49
30)	McClellan, Bernstein, & Garbin, 1984	26 psychology students	128 experimental stimuli	5	Magnitude estimations of fins-in and fins-out Mueller Lyer stimuli	Actual magnitude of fins-in and fins-out Mueller Lyer stimuli	Perception	r_0 = .72
31)	Trailer & Morgan, 2004	75 students	50 situations in a questionnaire	11	Predicting the motion of objects	Actual motion	Intuitive physics	r_0 = .15 (*, +)

Note. ▲= subjective criterion. r_0 = type of correlation is unknown. (*) = idiographic approach (cumulating across individuals). (*, +) = both research approaches are considered.

4.4 Meta-analysis: Cumulating research findings

The goal of science is to cumulate knowledge in theories. But, to begin with, scientists need an overview of the data. Before the development of a meta-analysis, narrative literature reviews provide an overview of the data in an area, and finally lead to a theory. Such literature reviews often show conflicting findings. Some studies, for example, find a statistically significant relationship between two variables of interest, while others do not report this fact (for details, see below). The main difference between literature reviews and the further development of a meta-analysis is that literature reviews are based on studies without cumulating them. Hence, the term meta-analysis "has become encompass all of the methods and techniques of quantitative research synthesis" (Lipsey & Wilson, 1993, p. 1). Glass (1976) summarised the term meta-analysis as follows:

> The term is a bit grand, but it is precise, and apt, and in the spirit of "meta-mathematics", "meta-psychology", and "meta-evaluation". Meta-analysis refers to the analysis of analysis. I use it to refer to the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings. (p. 3)

However, this difference between narrative reviews and meta-analysis leads to critique for instance by Hunter and Schmidt (2004) who claimed that "the perfect study is a myth" (p. 17). "There are no perfect studies" (p. 18), and therefore, such narrative literature reviews cannot answer any questions, because they are based on imperfect studies. Hence, Hunter and Schmidt suggested a meta-analytical approach that may provide a solution for correcting imperfect studies and allows the researcher to synthesize the data from multiple studies.

To summarize, one can say that in contrast to narrative reviews, a meta-analysis is a systematic and objective alternative for synthesizing empirical evidence. As this procedure requires informed judgment by the meta-analyst, however, methodologists still develop guidelines to conduct and report meta-analyses in order to increase the objectivity of the meta-analytic approach.

In the following we will give a historical overview starting from the first meta-analysis and concluding with the current spread of meta-analyses. Then, the weaknesses of the meta-analysis will be illustrated, and strategies to overcome them will be presented. As there are different meta-analytic approaches, we will focus on evaluation research, which led us to prefer the Hunter-Schmidt approach. Consequently, the Hunter-Schmidt approach is introduced in more detail. Finally, we will describe the methods of detection of publication bias.

4.4.1 Historical review

In the following, some of the numerous antecedents of the meta-analysis are described:

The first qualitative synthesis of findings from different studies was conducted by Pearson in 1904. He averaged the correlations between the inoculation for typhoid fever and mortality for five separate samples (see Cooper & Hedges, 1994).

An extension of Pearson's work is the early quantitative-review method well-known as the box-score approach by Meehl (1954, see also chapter 2.5.2). This work also introduces the "Statistical vs. Clinical Prediction" problem in psychology. Meehl's review includes 20 studies, in which predictions by clinical psychologists are compared with those of simple actuarial tables. In this case the success of clinical or actuarial predictions was marked and led to a frequency overview of these two types of predictions. However, clinical psychologists are usually outperformed by the actuarial predictions.

The narrative review on the effect of psychotherapy by Eysenck (1952) is also worth mentioning as an antecedent of the first meta-analysis method. In this review Eysenck's conclusion that psychotherapy has no beneficial effects on patients must have been a provocation for Glass and his experience as a therapist, which finally led him to a statistical evaluation of Eysenck’s conclusion. In 1970, Glass published his meta-analysis, which aggregated the findings of 375 psychotherapy outcomes and concluded that psychotherapy does indeed work (Smith & Glass, 1977).

At the same time, a contrary meta-analytic approach was developed by Hunter and Schmidt (see Hunter, Schmidt & Jackson, 1982).

In summary, both approaches, the Glass and the Hunter-Schmidt method, are to be considered starting points for the success of the meta-analysis as an influential research tool.

4.4.2 Actual spread of the meta-analysis

Since the late 1970s, when Glass and Hunter-Schmidt independently developed two different meta-analytic approaches, the power of meta-analyses increased. In the following, the spread of the meta-analysis is highlighted by the number of publications leading to efforts for study registration and to special issues on the topic.

The increasing success of the meta-analysis as a research tool is also clearly shown by Hunter and Schmidt’s internet search with the term “meta-analysis”. This search yielded 2’500 hits in 1999 and 552’000 hits in 2004 (see 2004, p. 24). Our repeated internet search with google generated 4’320’000 hits, although only 930’000 hits were indicated by the online search engine google scholar in 2007 (June). A year later, 6’260’000 hits resulted with a google and 1’210’000 with a google scholar search. Some authors use terms like “research synthesis” or “review” instead of the term “meta-analysis”. Therefore, these estimations clearly understate the actual spread of the meta-analysis method. However, these results showing a steady increase of the use of the meta-analysis are also in line with Schulze’s search (2007).

After a number of meta-analyses were published, some meta-analyses were rerun. So, also the meta-analysis by Glass was confirmed focusing on a German sample to find any psychotherapy effects (see Wittmann & Matt, 1986). In addition, the increasing use of the meta-analysis leads not only to replicated meta-analyses, but also to an aggregation of 350 meta-analyses published on the subject of psychological, pedagogical, and behavioural intervention in the early 1990’s (Lipsey & Wilson, 1993).

Relating to the increased use and number of publications on meta-analyses, efforts concerning the registration of studies are also established. The aim of study registration was to have easy access to all unpublished and published studies on a subject and to prevent publication bias (see chapter 4.6). The following three study databases are well-known (this list is chronologically ordered):

1) The *Cochrane Collaboration* was founded first and is a database of controlled clinical trails and systematic reviews (see http:// www.cochrane.org/). If a new study is available on this site, updated results are always available via the internet.
2) The *Campbell Collaboration* is a database for social sciences founded in 1999 (see http://www.campbellcollaboration.org/). It is supported by well-known researchers in the field, such as Borenstein, Hedges, Shadish.
3) The *What Works Clearinghouse* is a database for educational studies and reviews founded in 2002 by the U.S. Department of Education’s Institute of Education Sciences (see http://ies.ed.gov/ncee/wwc/).

The importance of meta-analysis in psychology also becomes visible in the special issues on meta-analysis in the *Journal of Psychology* (Zeitschrift für Psychologie) in 2007 and in *Organizational Research Methods* in 2008, although a special journal for meta-analysis does not exist in 2009, but a new journal, *Research Synthesis Methods*, has been launched.

In sum, the wide use of meta-analyses started in 1970, and more and more scientists are applying them to contrary findings in their research areas. About the actual situation of the meta-analysis as a research tool, Schulze (2007) concludes:

> Meta-analysis has earned its place in the pantheon of scientific methods. It became a standard method of research synthesis in many empirical research fields, especially in the social sciences. (p. 87)

4.4.3 The weaknesses of meta-analysis

Although meta-analysis is highly appreciated by researchers, some disadvantages should be acknowledged. In the following, four main disadvantages with recommendations of how to control them will be presented (see also Eysenck, 1994, and Table 7):

1) Publication bias or the "*File-Drawer-Problem*" is defined as a bias towards studies with significant results; they are more likely to be accepted. In our case, it could be that studies with positive correlations are more likely to be accepted for publication than studies with a negative correlation. This fact could be responsible for a considerable threat to the representativeness of meta-analysis samples (see chapter 4.6). In addition, one could assume that studies showing that experts are not as accurate in their judgments as students may have problems getting published. To prevent that any publication bias will influence the interpretation of our data, we considered the following strategies:
    a. We used a comprehensive literature-search strategy to decrease the possibility of overlooking studies. Hence, we also included data bases like *ERIC* (see Appendix B: Table 1) which include references to unpublished reports or conference papers.
    b. We checked the possibility of publication bias by means of graphics, so-called funnel plots (see chapter 4.6.1).
    c. We estimated the publication bias (see chapter 4.6.2) with different estimators, in order to find out how many studies are needed to change the actual results.

2) “Apples and oranges problem” represents the fact that in meta-analysis, studies that do not really deal with the same constructs and relationships are often integrated and summarized. Consequently, we carefully coded the studies to reveal any uniformity problems in our meta-analysis. Furthermore, we will consider this issue in our robustness analysis (see chapter 5.2.4 and Kaufmann & Athanasou, 2009).

3) The focus on *the quantitative approach* may lead to a negligence of the qualitative approach of the reviews. To not overlook the quality of the included studies, we also included the approach suggested by Slavin (1986), the so-called best-evidence synthesis, an attempt to combine qualitative and quantitative reviewing techniques in the same research review.

4) “*Garbage in – garbage out*” problem: This represents the fact that studies of different methodological quality are included. Slavin (1986) suggests to define very strict methodological criteria for inclusion, and so the meta-analyst has assurance that the synthesis is based on only the “best” evidence. Consequently, we will focus on the inclusion criteria and consider this fact in the coding of our studies and in our robustness analysis.

Table 7

Summary of disadvantages of meta-analysis and our suggested solutions

Disadvantages	Solution
Publication bias	- Comprehensive literature search (see chapter 4.1) - Funnel plots (see chapter 4.6.1) - File-safe-N calculation (see chapter 4.6.2)
“Apples and oranges problem”	- Coding (see chapter 4.2) - Robustness analysis (see chapter 5.2.4)
Quantitative aspects	- Evidence synthesis approach (see chapter 4.3)
Garbage in – garbage out	- Coding (see chapter 4.2) - Inclusion criteria (see chapter 4.1.4) - Robustness analysis (see chapter 5.2.4)

4.4.4 Different meta-analysis approaches

Over the past 30 years, a number of variants of the meta-analysis were developed, such as the Hedges-Olkin (1985), the Rosenthal-Rubin (see Rosenthal, 1991), and the Hunter-Schmidt (2004) method of meta-analysis (for an overview see Bangert-Downs, 1986; Rosenthal & DiMatteo, 2001).

As can be seen in Table 8, they are different in several points, such as *effect size*, applied *model*, or as to whether they use a *correction* procedure, and, finally, in what type of *test* they apply to identify any possible moderator variables.

Table 8

Methodological characteristics of the three dominant variants of meta-analysis: the Hedges-Olkin (1985), the Rosenthal-Rubin (see Rosenthal, 1991), and the Hunter-Schmidt (2004) approach

Methodological		Meta-analytic approaches	
Characteristics:	Hedges-Olkin (1985)	Rosenthal-Rubin (1991)	Hunter-Schmidt (2004)
Effect size	d	d	r
Model	fixed-effect model	fixed-effect model	random-effect model
Correction	no correction	--	artefact corrections
Test	Q	Q	75% rule

First, as you can see in Table 8, *effect size* belongs to two families: the r, correlation, and the d family (see Rosenthal, 1991). The d family comprises standardised mean differences and is available of studies reporting the results of experiments. On the other hand, in the r family, the correlation coefficient describes a bivariate relationship. However, one key feature of meta-analysis is the conversion of effect sizes. Hence, this meta-analysis characteristic is neglectable.

Secondly, you can also see that in meta-analytic research two different models are used; the fixed-effects and the random-effects model. The two models have different assumptions regarding the underlying population. A fixed-effect model assumes that all of the studies in the meta-analysis are derived of the same population and that the true size of the effect will be the same for all of the studies in the meta-analysis. Hence, the source of variation in the effect size is assumed to be variations within each study, such as for instance sampling error. In contrast to the commonly used fixed-effects model Hunter and Schmidt (2004) recommend a random-effects approach. The random-effects model assumes that population effects vary from study to study. The idea behind this is that the observed studies are samples drawn from a universe of studies. Random-effect models have two sources of variation in a given effect size: that arising from within the study itself and its (the source) from variations in the population effect between studies. However, the variation of effects from study to study appears to be the rule rather than the exception for most-real-world data. Consequently, the random-effects model seems to be more adequate for our analysis (see also Kisamore & Brannik, 2008, p. 52). It should be noted, however, that assumptions made by random-effects models are more tenable, in general, than those made by fixed-effects models, although most of the meta-analyses published in *Psychological Bulletin* are based on fixed-effects models (Kisamore & Brannick, 2008). There are also exceptions using random-effects models (see Karelaia & Hogarth, 2008).

Thirdly, we would like to mention that most methods of meta-analysis are concerned with only one *correction* strategy, namely the artifactual source of variation across studies, the so-called sampling error. The Hunter-Schmidt method is the only method which allows to correct studies for 10 further artefacts, such as, for example, measurement error (see Hunter & Schmidt, 2004, p. 18, for an overview see Table 10).

Finally, the last mentioned meta-analysis characteristics the used test (i.e. Q test or 75% rule) for identify any moderator variables are presented (see chapter 4.5.1.3).

4.4.5 Evaluation research on meta-analysis approaches

Although the approaches are different, there are also studies that compare and evaluate them. In the following, we will introduce the recent evaluation research on meta-analysis in more detail (see Table 9).

Field (2001) conducted two Monte Carlo studies to compare three meta-analytic approaches. This study shows that in the most common case in meta-analytic practice the Hunter-Schmidt method tends to provide the most accurate estimates of the mean population effect size (see also Hall & Brannick, 2002; Field, 2001). Beside these simulation studies, also studies on real data support the use of the Hunter-Schmidt method (see Kisamore & Brannick, 2008).

Further research on the comparison of meta-analytic procedures shows that the Hunter-Schmidt method is more precise than the Hedges-Olkin approach when it comes to point estimates, homogeneity tests[5] to prevent Type I error rates, the error of rejecting a hypothesis when it actually should be accepted (see Aguinis, Sturman, & Pierce, 2008). However, this analysis is based solely on simulations. Studies based on real data are not available on this subject at the moment.

Consequently, we can summarise the introduced evaluation research on meta-analytic procedure to the effect that the Hunter-Schmidt method is more precise than the Hedges-Olkin method – but also more conservative. In addition, our selection of the Hunter-Schmidt approach is also supported by the fact that the mentioned *LME* (Tucker, 1964) is the base for the Hunter-Schmidt approach (for more details, see chapter 4.5.2.1).

[5] Although the Hunter-Schmidt method does not advocate the use of null hypothesis significance testing, a statistical significance test was performed.

Table 9

Summary of the current evaluation research on meta-analytic approaches

Studies:	Investigation:	Results:
Field (2001)	model	random-effect model
Kisamore & Brannick (2008)	model	random-effect model
Aguinis et al. (2008)	*Performance:*	Hunter-Schmidt
	Point estimates	
	Homogeneity tests:	
	Type I error rates	Hunter-Schmidt
	Type II error rates	Hedges and Olkin
	Moderating effect tests:	
	Type I error rates	Both
	Type II error rates	Both

4.5 Hunter-Schmidt approach

As mentioned before, our analyses follow the steps recommended by Hunter-Schmidt (2004). Hunter and Schmidt's interest in the differential validity of employment tests for blacks and whites (Schmidt, Berner, & Hunter, 1973) led them to develop a quantitative research-synthesis tool for this area. Besides its most extensive use in the domain of personnel testing (see Hunter et al., 1982), it is also applicable for the assessment of the validity of any measurement procedure. In the beginning, this method was called validity generalization, because the original goal was to develop a research tool to estimate the population value (i.e. true value, validity value). With this method, the validity of one study can now be inferred from the validity found in hundreds of previous studies. This meta-analysis procedure determines the degree to which validity findings can be generalized. These days, the Hunter-Schmidt method indicates that all or most of the study-to-study variability due to artefacts and the traditional belief in personal selection of a situation-specific validity of tests was erroneous (Hunter & Schmidt, 2004, p. 160).

However, the purpose of conducting a meta-analysis according to Hunter and Schmidt (2004) was to determine whether the variance in reported *LME* components was entirely the result of statistical artefacts. We would like to mention that such artefacts are often falsely interpreted as conflicting findings in reviews – instead of sampling error – and therefore lead to wrong conclusions. However, Hunter et al. (1982) have recommended that research integrators correct their correlation coefficients and the associated variances for statistical artefacts (like sampling or measurement error). Hence, it is unique for this meta-analytic approach that there are two types of meta-analysis: the bare-bones meta-analysis and its extension, the psychometric meta-analysis. A bare-bones meta-analysis is only corrected for sampling error. A psychometric meta-analysis is also corrected for other artefacts.

Furthermore, the main difference between the Hunter-Schmidt method and the latter is in the use of untransformed correlation coefficients instead of Fisher's z transformation in the correcting procedure.

Finally, it must be mentioned that in our data base sometimes only the data of individuals (idiographic approach) is reported. In this case, the Hunter-Schmidt method is used, but across persons, using individual as analysis unit. This type of within cumulating is symbolized by a (*) in Tables 5 and 6 in the last column. In the following, we will therefore illustrate the two types of meta-analysis – firstly describing the use of individual data, and then using data across individuals.

4.5.1 Bare-bones meta-analysis

4.5.1.1 Idiographic data base

To overcome the weakness of ecological fallacy, we tried to obtain individual data from as many studies as possible and to control our analysis for ecological fallacy with this data base. Therefore, we also used the idiographic research approach; In this case, r_i is a component of correlation of the *LME* (e.g. the achievement correlation) of person *i*, and N_i is the number of judgments of person *i* (e.g. 178 forecast days, see Table 6). It is to mention that this weighting strategy is different from that suggested by Hunter and Schmidt (2004). Hence, we will check this weighting strategy in our robustness analysis (see chapter 5.2.4.2)

Furthermore, since sampling error cancels out in the average correlation across studies, we estimated the mean population correlation ($\bar{r}$, see Equation 2, Hunter & Schmidt, 2004, p. 81) in our meta-analysis by means of the sample correlations.

$$\bar{r} = \frac{\Sigma[N_i r_i]}{\Sigma\ N_i} \qquad (2)$$

However, sampling error adds to the variance of correlations across persons. Therefore, the observed variance (σ_r^2*,* see Equation 3, Hunter & Schmidt, 2004, p. 81) is corrected by subtracting the sampling error variance (σ_e^2*,* see Equation 4, Hunter & Schmidt, 2004, p. 89). The resulting difference is then the variance of population correlation across persons.

$$\sigma_r^2 = \frac{\sum\left[N_i\left(r_i - \bar{r}_i\right)^2\right]}{\sum N_i} \qquad (3)$$

$$\sigma_e^2 = \frac{\left(1 - \bar{r}^2\right)^2}{\left(\bar{N} - 1\right)} \tag{4}$$

Furthermore, the average sample size ($\bar{N}$) was calculated as follows (see Equation 5, Hunter & Schmidt, 2004, p. 88):

$$\bar{N} = T / k \tag{5}$$

where *T* is the total number of judgments across persons, and *k* is the number of analyzed judgments (e.g. 370 for the number of achievement analyzed judgments across studies, see chapter 5.1).

Furthermore, in meta-analysis according to Hunter and Schmidt (2004, p. 205), credibility and confidence intervals are distincted. In contrast to the used confidence intervals, credibility intervals do not depend on sample size, and, hence, sampling error. Therefore, a credibility interval is an estimate of the range of real differences after accounting for the fact that sampling error may be due to some of the observed differences. If the lower credibility value is greater than zero, one can be confident that a relationship generalizes across persons examined in the study. As Hunter and Schmidt (2004) concluded that: “credibility intervals are usually more critical and important than confidence intervals” (p. 206), we used 80% credibility intervals in our analysis, formed by SD_ρ as follows (see Equation 6):

$$\bar{\rho} = \pm\ 1.28 * SD_\rho \tag{6}$$

4.5.1.2 Nomothetic data base

According to Hunter and Schmidt (2004, p. 442, see also Athanasou & Cooksey, 1993), subgroups (i.e. judgment tasks) of the total study correlation are used for the meta-analysis across judgment tasks. Subgroup correlations are symbolized by a roman numeral in Tables 5 and 6. To summarize: Our included 31 studies are separated into 49 different judgment tasks. Hence, we used the described Hunter-Schmidt method with the equations 2 to 6 also for this meta-analysis, but across judgment tasks.

4.5.1.3 Moderator variables

To detect moderator variables, we focused on assessment with the 75% rule (see Sackett, Harris, & Orr, 1986). As mentioned before, Hunter and Schmidt suggested subtracting the variation due to sampling error from the total variation. If sampling error removes approximately 75% of the overall variation, they conclude that the effect sizes are homogeneous, due the fact that they estimate one parameter.

However, if the 75% rule indicates a lack of homogeneity of a single effect sizes, a search for a moderating variable is conducted. A variable *Z* (e.g. the applied research area) is a moderator variable of the relationship between variables *X* (e.g. a judgment) and *Y* (e.g. actual outcome), when the nature of this relationship is contingent upon values or levels of *Z*. Research approach, research area, and experience level within research area are candidates for moderator variables in the presented meta-analysis (see also chapter 3). The data set is then split up according to the categories of the moderator variable, and separate meta-analyses are performed on each subset of data. It should be mentioned that moderator analyses are by nature observational studies, i.e. the meta-analyst simply observes, in retrospect, the characteristics of the studies (such as the research area). Therefore, the results from a moderator analysis do not provide any evidence of a causal relationship between variables *Z* and *Y*. Furthermore, a spurious relationship between variable *Z* and *Y* could be introduced by a moderator analysis.

4.5.2 Psychometric meta-analysis

In contrast to other meta-analysis methods, the Hunter-Schmidt approach is the only one that allows the correction of 11 artefacts. This psychometric approach estimates the population correlation by correcting the observed correlations for downward bias due to various artefacts (see Hunter & Schmidt, 2004, p. 35). However, the Hunter-Schmidt approach bases on the assumption that the perfect study is a myth (see Hunter & Schmidt, 2004, p. 17). This assumption is in line with Rubin (1990) as follows:

> Under this view, we really do not care scientifically about summarizing this finite population (of observed studies). We really care about the underlying scientific process – the underlying process that is generating these outcomes that happen to see – that we, as fallible researchers, are trying to glimpse through the opaque window of imperfect studies. (p. 157)

Finally, an overview of all suggested artefacts by Hunter and Schmidt (2004, p. 35) leads to an approximation of an accuracy estimation based on imperfect studies; the suggested artefacts are listed and described by an example in the following Table 10. To summarize: Artefacts are sample bias, measurement error, or bias such as dichotomization of continuous dependent and independent variables, deviations from perfect construct validity in the dependent and independent variables, transient errors of measurement, and, finally, random response of errors of measurement, measurement error due to scorer disagreement, and variance due to extraneous factors.

Table 10

Description of 11 artefacts that alter the value of outcome measures according to Hunter and Schmidt (2004, p. 35), with the study by Cooksey et al. (1986) as an example

1. Sampling error:
   E.g.: Study validity will vary randomly from the population value because of sampling error.

2. Error of measurement in the dependent variable:
   E.g.: Study validity will be systematically lower than true validity to the extent that a teacher's reading-achievement estimation is measured with random error.

3. Error of measurement in the independent variable:
   E.g.: Study validity for a standardized test score (criterion) will systematically understate the validity of the actual reading achievement measured, because the actual standardized test score is not perfectly reliable.

4. Dichotomization of a continuous dependent variable:
   E.g.: The teacher's reading-achievement estimation could artificially be dichotomized into "successful" or "not successful", although the estimate was in the form of a percentage score with possible values ranging form 0% to 100%.

5. Dichotomization of a continuous independent variable:
   E.g.: The actual standardized test score could be artificially dichotomized into "successful" versus "not successful".

6. Range variation in the independent variable:
   E.g.: Study validity will be systematically lower than true validity to the extent that the teacher's reading-achievement estimation causes students to have a lower variation in the actual test score (criterion) than is true.

7. Attrition artefacts: Range variation in the dependent variable:
   E.g.: Study validity will be systematically lower than true validity to the extent that there is systematic attrition in students' reading achievement, e.g. when good students are promoted out of the population, or when poor students are shut out from this class due to poor achievements.

8. Deviation from perfect construct validity in the independent variable:
   E.g.: Study validity will vary if the factor structure of the reading test differs from the usual structure of reading tests for the same trait.

9. Deviation form perfect construct validity in the dependent variable:
   E.g.: Study validity will differ from true validity if the actual reading achievement (criterion) is deficient or contaminated.

10. Reporting or transcription error:
    E.g.: Reported study validity differs from actual study validity due to a variety of reporting problems: inaccuracy in coding data, computational errors, errors in reading computer output, typographical errors by secretaries or by printers.
    Note: These errors can be very large in magnitude.

11. Variance due to extraneous factors that affect the relationship:
    E.g.: Study validity will be systematically lower than true validity if students differ in reading achievement at the time their performance is measured (because reading experience affects reading achievement).

4.5.2.1 An extension of Tucker's Lens Model Equation

As mentioned above, there is a relation between Tucker's *LME* (1964) and the meta-analytic approach according to Hunter and Schmidt (2004), although they not refer to it. However, there is a historical connection, as Tucker supervised Schmidt's thesis. The corrected judgment achievement in our example can furthermore be estimated empirically according to Hunter and Schmidt (2004) and its extension by Wittmann (1988) as follows:

$$r_{a,\ true\ value} = \underbrace{S}_{\underbrace{\substack{\text{Selection effects}\\ \text{due to restriction}\\ \text{(enhancement)}\\ \text{of range}}}_{\substack{\text{1 Danger to}\\ \text{overestimate}\\ \\ \text{1 Danger to}\\ \text{underestimate}}}} \sqrt{\underbrace{r_{tt}^{Rs}\, r_{tt}^{Re}}_{\underbrace{\substack{\text{Psychometric}\\ \text{reliability of}\\ \text{judgment and}\\ \text{criterion}}}_{\substack{\text{2 Dangers to}\\ \text{underestimate}}}}} \qquad G\, \underbrace{R_s\, R_e}_{\underbrace{\substack{\text{Environmental}\\ \text{validity and}\\ \text{consistency}\\ \text{(Construct}\\ \text{reliability)}}}_{\substack{\text{2 Dangers to}\\ \text{underestimate}\\ \text{(lack of}\\ \text{symmetry)}}}} \quad + \quad \underbrace{e}_{\underbrace{\text{Sampling error}}_{\substack{\text{1 Danger to}\\ \text{overestimate}\\ \text{(positive error)}\\ \text{1 Danger to}\\ \text{underestimate}\\ \text{(negative error)}}}} \tag{7}$$

Researchers interested in Brunswik research know that the famous *LME* is traced to Brunswik. As you can see from Equation 7, the linear part (i.e. GR_sR_e) of the *LME* is one part of our meta-analyzed judgment achievement estimation. This part is multiplied by psychometric concepts.

Finally, the sampling error is added. As mentioned above, a bare-bones meta-analysis is only corrected for sampling error. In a sampling-error correction, there is a danger to overestimate the true correlation value (judgment achievement), leading to a positive error. On the other hand, there is also the danger of underestimating judgment achievement, a so-called negative error.

A psychometric meta-analysis, however, includes more artefact corrections than only sampling error (see Table 10). In Equation 7 you will find artefact corrections like as reliability, validity, and selection effect. Although Hunter and Schmidt (2004) recommend 11 corrections for artefacts, they ignored the symmetry concept. The symmetry principle implies that judgment achievement is only maximal, if the judgment is made on the same level as the criterion. Otherwise, judgment achievement is not optimal. According to Wittmann (1985, 1988), there are four violations against symmetry.

In our presented work we do not consider the symmetry concept; this should urgently be done in further research. Therefore, we concluded that our meta-analysis will underestimate the actual value, unless the symmetry concept is considered.

In summary, in Equation 7 it is visible that a psychometric meta-analysis leads to six dangers of underestimating to two dangers of overestimating the true judgment achievement value. In the following, we therefore use a psychometric meta-analysis to estimate judgment achievement as accurately as possible.

4.5.2.2 Procedure

Artefact information is not always available from our studies. In our example, we get sample size information from all studies. However, the other artefacts (such as the reported reliability) in studies are only sporadically available. As missing data for correcting artefacts is common in meta-analysis studies, Hunter and Schmidt (2004, p. 137) propose correction by means of distribution of artefact values, which is complied across the studies that provide information on that artefact. Therefore, we used the method of artefact distribution. Consequently, we conducted a meta-analysis in two stages: A bare-bones meta-analysis corrects for those artefacts for which information is available for all studies, in our case only for sampling error. Secondly, we estimated the artefact distribution on the available information in a psychometric meta-analysis.

As the first step a bare-bones meta-analysis is already introduced, we will focus on a psychometric meta-analysis in more detail based on our idiographic data base before we report the psychometric meta-analysis used with our nomothetic data base.

4.5.2.2.1 Idiographic data base

According to the mentioned introduction also in the psychometric meta-analysis procedure for idiographic studies, each person is treated as a single study. Therefore, to keep our methodological introduction short, we refer to the chapter 4.5.2.2.2, which explains in more detail a psychometric meta-analysis applied to studies with a nomothetic research approach. This description can also be applied to the idiographic approach in that each study refers to a single person.

4.5.2.2.2 Nomothetic data base

As mentioned before, the psychometric meta-analysis bases on a bare-bones meta-analysis. This procedure has already been explained (see chapter 4.5.1), and, we will therefore only mention the supplemented steps for a psychometric meta-analysis (Hunter & Schmidt, 2004, p. 181) and the additional artefact corrections in the following.

4.5.2.3 Artefacts

According to the available data, we can only consider two artefacts in our psychometric meta-analysis: measurement error and dichotomization.

4.5.2.3.1 Measurement error

Because decision and judgments measurements are not always without error, the reliability values should also be considered, in order to find out how well the validity of judgement and decision making actually is. The reliability is therefore always the basis for validity $_{(\max)}r_{tc} = \sqrt{r_{tt}}$. Reliability is defined as the correlation between parallel tests and interprets this reliability as the ratio of true-score variance to observed-score variance (see Wiggins, 1973, p. 282). According to Wiggins (1973, p. 283, see APA, 1954, p. 28), "reliability is a generic term referring to many types of evidence". Furthermore, Wiggins (1973) mentions that:

> Clearly, different designs for determining the reliability of parallel observations take account of quite different sources of error. Thus, although reliability may be defined as the ratio of true-score variance to observed-score variance, the error that enters into observed scores differs from one design to another. Internal-consistency procedures involve the estimation of error due to the selection of a given set of items or observations. Depending on the time interval between administrations of parallel forms, equivalence procedures may estimate error due to selection of specific items and/or to response variability of subjects. Stability procedures provide an estimate of response variability in subjects as well as of the effect of differences in conditions of test administration or observation. (p. 283)

However, as mentioned before, variables in science are never perfect measures (for an overview, see Schmidt & Hunter, 1996). This leads to error of measurement and systematically lowers the correlation between measures in comparison to the correlation between the variables themselves. Reliability coefficients represent the measurement error in each study. In our case, we had to correct judgments and criteria' (see Figure 3) for measurement error. Hence, we will first introduce our measurement corrections in judgments, and then on the criteria side.

An overview of the included studies shows that only three studies reported reliability coefficients. The correlation coefficient for each person is reported in the studies by Levi (1989, r = .73 - .93) and Athanasou and Cooksey (2001, r = .20 - .99). Athanasou and Cooksey (2001) calculated the retest reliability by selecting 20 random scenarios out of 100 scenarios and then adding them to the 100 scenarios as a repeated task. The study by Wiggins and Kohen (1971, r = .09) reports an aggregated reliability coefficient.

For the missing retest-reliability information, we used the review on “Test-Retest Reliability of Professional Judgment” by Ashton (2000) to estimate judgments corrected for measurement error. An advantage of this review is its separation into different research areas, such as medical science and business science. Taking medical science as an example, we used the mean of the test-retest reliability of .73 (.76 for medical doctors; .70 for clinical psychologists) to correct the judgments for measurement error. In addition, we used the retest-reliability values for meteorologists' hail forecasts (.93, see Ashton, 2000) for all meteorologist forecasts in our analysis.

As mentioned before, also the measurement error in the criterion variable is considered. We defined three types of criteria: objective, subjective, and test criteria. The criterion is measured as objective for example if a physiologic measurement of the patient's actual hemodynamic status (see Speroff et al., 1989, see Table 5) is used for a criterion. Consequently, the test-retest reliabilities of our criteria were corrected with the value 1 for objective criteria. We therefore entered 1 into our data base for the reliability of the predictor, because we did not correct for measurement error, assuming that machine measurement is 100% correct. However, in psychological tests or tests not measured by a machine, the test criteria values were corrected by other test-retest reliabilities by specific tests, such as the *MMPI* (see Einhorn, first study, 1974, r_{tt} = .71, see Nunnally & Bernstein, 1994) or the *Wonderlic Personnel Test* (see Reynolds & Gifford, 2001, r_{tt} = .94, see Dodrill, 1983). Finally, if a subjective value like the judgment of a single physician (see LaDuca et al., 1988, Table 5) is used, also the values of Ashton's review (2000) are applied to correct the measurement error (r_{tt} = .76 for medical doctors). In Table 5, all subjective criteria are marked with a triangle in the criterion column.

Finally, it is to mention that because of missing data we mostly used aggregated retest-reliability values for our meta-analysis.

4.5.2.3.2 Dichotomization

In the following, the dichotomization of a continuous variable is considered. Many decisions, such as medical decisions (healthy or diseased) or job application decisions (accepted or not accepted), are binary. It should now be considered, that often such decisions are based on continuous criteria – like scores of medical tests that are dichotomized by using a cut-off value. So, "if a continuous variable is dichotomized, the point-biserial correlation for the new dichotomized variable will be less than the correlation for the continuous variable" (see Hunter & Schmidt, 2004, p. 36). This artificial dichotomization may lead to an underestimation of the validity.

An overview of our studies shows that only the study by Szucko and Kleinmuntz (1981) uses a point-biserial correlation. It can not be excluded that other studies with unknown types of correlation coefficients include further point-biserial correlations.

According to Hunter and Schmidt (2004, p. 36) we used the correction formula of a double dichotomization (see Equation 8):

$$\rho_0 = a\rho \tag{8}$$

where a = .80 (see Hunter and Schmidt, 2004, p. 36). Consequently, the point-biserale correlation of .23 increases 20%, so, the corrected correlation used in our meta-analysis for the Szucko and Kleinmuntz (1981) study is actually estimated as .27 based on nomothetic data. In the Appendix E: Table 1 you will also find the corrected single judges' values used for our meta-analysis based on individual data.

4.5.2.4 Corrections of artefact information

For the detailed explanation of our artefact corrections we refer to the Appendix E. To summarize: We used the following three steps recommended by Hunter and Schmidt (2004):

1) Cumulation of artefacts information
2) Correction of the mean correlation
3) Correction of the standard deviation of correlations.

It is important to note that in the following psychometric procedures the estimation of 80% credibility interval, the 75% rule, and, finally, the detection of moderator variables is the same as in a bare-bones meta-analysis (see chapter 4.5.1). Consequently, the same steps as already reported are applied.

4.6 Publication bias

4.6.1 Funnel plots

As publication bias of the included studies is considered (see also chapter 4.4.3), a funnel plot (Light & Pillemer, 1984) evaluating the extent of the publication bias is illustrated. The funnel plot for all correlations of judgment achievement in the 49 judgment tasks included in our meta-analysis is presented in Figure 8.

The plot should look like a funnel (see dashed lines), when sample size is plotted on the x-axis and achievement correlations on the y-axis, because small samples are expected to show more variability than large samples. A not perfect funnel plot is yielded. To check for publication bias, the trim-and-fill method suggested by Duval and Tweedie (2000) was used to estimate the missing studies (see red triangles in Figure 8). Hence, in our robustness analysis we estimated the missing studies and supplemented our data base with them assuming only objective criterions in a psychometric meta-analysis (see chapter 5.2.2) before rerunning our analysis.

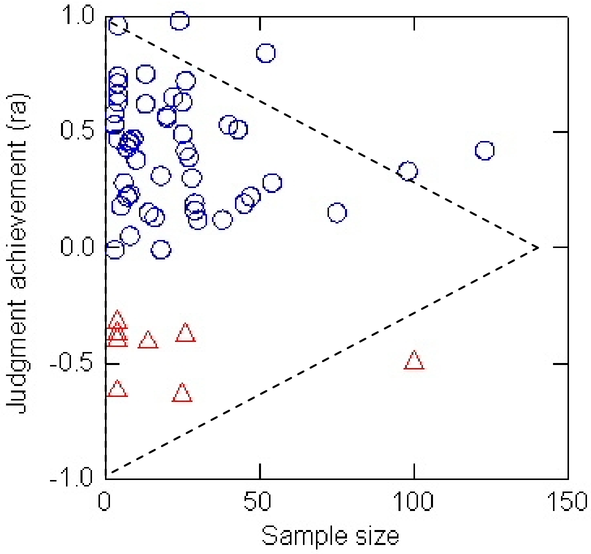


Figure 8. Funnel plot of achievement correlations *(r_a)* versus sample size for the 49 tasks included in our meta-analysis.

4.6.2 Calculating Fail-safe numbers

In the following analysis, the same sample, judgment achievement of the included tasks in our meta-analysis, is used for the estimation of the Fail-safe number suggested by Orwin (1983). This Fail-safe number indicates the number of no significant, unpublished (or missing judgment achievement tasks) studies that would need to be added to a meta-analysis in order reduce an overall statistically significant observed result to no significance. If this number is large relative to the number of observed studies, one can feel fairly confident in the summary conclusions. Rosenthal (1979) suggested the “five plus ten rule”, which means that if the Fail-safe number is not more than five times the number of reviewed studies plus ten, the obtained findings are probably robust.

The Fail-safe numbers were calculated with an SPSS (2004) syntax[6]. It must be mentioned that in the following analysis judgment tasks with three or less judges (see Einhorn, 1974, second study; Kim et al.,

[6] http://pages.infinit.net/rlevesqu/Syntax/MetaAnalysis/MetaAnalysisOfCorrCoef2.txt

1997) are excluded, which leads to on a slight overestimation of our results.

However, the Faile-safe number of 61 concerns publication bias, leading this meta-analysis to dramatically overestimate the achievement correlations (see Table 11). Our analysis shows clearly that according to the rule of thumb by Rosenthal (1979), all calculations have the tendency of publication bias. However, a closer look at the data reveals on the one hand that in the overall analysis 61 judgment tasks are needed to change the results; hence, as this is more than double the data base, we assume that there is no publication bias in all overall calculations for the *LME* components, except for component *C*. On the other hand, there is a clear publication bias in all *C* calculations as well as in all sub-analyses, which should be considered in the interpretation of our results and in our robustness analysis.

4.7 Calculations

All further calculations were done with the Hunter-Schmidt meta-analysis program (Schmidt & Le, 2005). In addition, for our publication and robustness analysis the program R (2007) was used.

Furthermore, the meta-analysis follows the Campbell Collaboration Guidelines (2007) and suggestions by Shadish (2007) and Egger, Smith and Altman (2001).

Table 11

Publication bias tendency according to Orwin's (1983) Faile-safe number

	Components				
Research area:	r_a	G	R_s	R_e	C
Medical science	9	19	23	29	0
Business science	16	20	21	22	- 4
Educational science	4	12	10	13	- 4
Psychological science	7	16	39	39	16
Miscellaneous	19	41	42	33	-10
Experience:					
Experts[a]	17	40	67	49	0
Business	4	8	10	13	-2
Education	4	7	5	7	-2
Psychology	-2	-1	12	14	-1
Miscellaneous[a]	[b]	[b]	[b]	[b]	[b]
Students[a]	32	62	58	67	-8
Business	10	11	10	10	-1
Education	1	4	5	5	-2
Psychology	3	14	28	25	14
Miscellaneous[a]	10	22	25	18	-6
Overall	61	122	139	118	- 8

[a] 4 judgment tasks were excluded, because they include only two persons (see Stewart et al., 1997). [b] was *not calculated* because the sample size was too small (i.e. 4 judgment tasks included with only two persons, see Stewart et al., 1997).

5 RESULTS

In the following, our results are presented at three different levels: First, we will focus the individual level without considering any meta-analysis, followed by our meta-analysis – first based on individual data and then on nomothetic data, both separated into a bare-bones and a psychometric meta-analysis.

Due to the fact that in some studies one component is missing, the sample sizes vary between the components. This may restrict our possibilities to interpret achievement in terms of relations between components within studies to a minor extent.

In our meta-analysis, the components of correlations (from -1.00 to 1.00) of the *LME* were interpreted according to Cohen's (1988) standards, with absolute values ≤ .29 considered as small, ~ .49 as moderate, and ≥ .50 as large magnitudes.

5.1 Idiographic data base

Before presenting our results, we would like to mention that a similar analysis has already been published (see Kaufmann et al., 2007). In contrast to our earlier analysis, the current analysis varies in these points:

a) We did not include four studies (Ashton, 1982; Lehman, 1992; Trailer & Morgan, 2004; Werner et al., 1989) because our previous literature search did not reveal them. Hence, also the number of single judgments analyzed by the *LME* has increased from 264 to 370.

b) In this analysis, we separated the combined category educational or psychological research area into two distinct categories. This categorisation is now in line with our meta-analysis based on nomothetic data.

c) In our current analysis, we added an analysis on the experience level within the different areas.

d) We also calculated missing component values (see Appendix C).

e) We would like to mention that we used another analysing tool (Hunter-Schmidt meta-analysis program, Schmidt & Le, 2005, instead of the SPSS syntax written by Marta Garcia-Granero and adapted by Wright, 2005).

f) Finally, we supplemented the already published bare-bones meta-analysis with a psychometric meta-analysis.

To summarise: The following presentation is a more elaborated analysis of our previous publication.

To begin with, we will overview the extreme values of judgment achievement. Consequently, three decision makers with a low judgment achievement and three decision makers with a high judgment achievement are described and compared in the following Table 12.

Table 12

Correlation components of three judges with high judgment achievement and three judges with low judgment achievement

Study	Components				
High judgment achievement	r_a	G	R_s	R_e	C
Stewart et al. (1997)	.97	.99	.98	.97	.46
LaDuca et al. (1988)	.75	.89	.88	.93	.17
Ashton (1982)	.88	.98	.96	.95	-.10
Low judgment achievement					
Szucko & Kleinmuntz (1981)	.02	-.17	.47	.52	.09
Wright (1979)	.27	.70	.62	.02	.34[a]
Trailer & Morgan (2004)	.14	.54	.26	.98	.00[a]

Note. A similar table was published in Kaufmann et al., 2007. We adapted this table to our actual analysis. [a]These values are not founded in publications, and we therefore calculated them by ourselves, see Appendix C.

The highest value of judgment achievement is found in a meteorological temperature forecast (Stewart et al., 1997, see Table 12). The components of the *LME* are large, reflecting an optimal decision condition. The task is highly predictable, and the meteorologist uses cues with high consistency. Judgment achievement is nearly optimal, because it is almost equal to the (linear) knowledge component. It is notable that this component is also the maximal value of all error-free judgment values across persons. A comparison of single judges with high judgment achievement shows that even the other components are high, with the exception of component *C*, which leads to a great variation from -.10 to .46 across different research areas (see Table 12).

To enhance our knowledge about the underlying sources of judgment achievement, we also took an interest in single judges with low judgment achievement. The lowest achievement value shows a correlation in the wrong direction (-.13, Einhorn, 1974, second study), i.e., greater judged severity does not match lower rates of survival. The physician was moderately consistent (.48), and also the task was moderately predictable (.30). The individual analysis of the *LME* shows that the physician in this study used information not explicitly available in the cues picked by a physician. However, that the underlying sources of poor judgment achievement can vary is apparent from the last three cases in Table 12. The low achievement level of the judge in the study by Szucko and Kleinmuntz (1981) indicates that if the judge could acquire better knowledge he would achieve better judgment, provided that the high consistency remains. In contrast, the low judgment achievement of a judge from the study by Wright (1979) indicates low task predictability, and therefore, poor knowledge or lack of consistency is not the reason for the mentioned low judgment achievement. The last case (Trailer & Morgan, 2004) shows that low judgment consistency can also be associated with poor achievement level.

From an idiographic point of view, it may be of interest to compare two studies with seemingly equal objective, concrete criteria. Two such studies are Einhorn (1974) and Stewart et al. (1997, see Tables 5 and 6), both including experts. The first study used “patients’ months of survival” as a criterion, the latter study “actual temperature” as criterion, in both studies thus an objective, concrete criterion. Despite this formal similarity between criteria, the studies present very different achievement values. In the study by Stewart et al. (1997) we found our highest achievement value

(.97), while the study by Einhorn presents a negative achievement value (-.13), and also our lowest judgment achievement value. Even though there may be several underlying factors responsible for this large difference in factors we are only able to speculate about, we can still pose a question: Are criteria generally regarded as equally objective or concrete also perceived in the same way by the single judge, i.e. as equally objective and concrete?

However, in a first step, the descriptive statistics applied to our data based to the 370 judgment achievements reveal that half of them (49%) are low, and 33% are high, and only 17% are medium (see Appendix F: Table 1). In addition, a similar pattern is found in the medical and in other research areas, but clearly not in educational science. In the educational area 69% of the included judgment achievements are high.

Finally, although the reported three judges in Table 12 with high judgment achievement are all experts, this should not imply that experts have better judgment achievement. If we compare judgment achievement across all areas by experienced and inexperienced judges (i.e. students), there is no tendency that experts reach a better judgment achievement at first glance.

5.1.1 Bare-bones meta-analysis

In the following, the meta-analytic results of the idiographic approach are presented in two sections. The first section describes the results for the achievement correlations across the judgment tasks presented in Figure 9 and Table 13. The second section reveals the additional *LME* components across the judgment tasks in Figures 10 to 12 and Table 14.

5.1.1.1 Judgment achievement

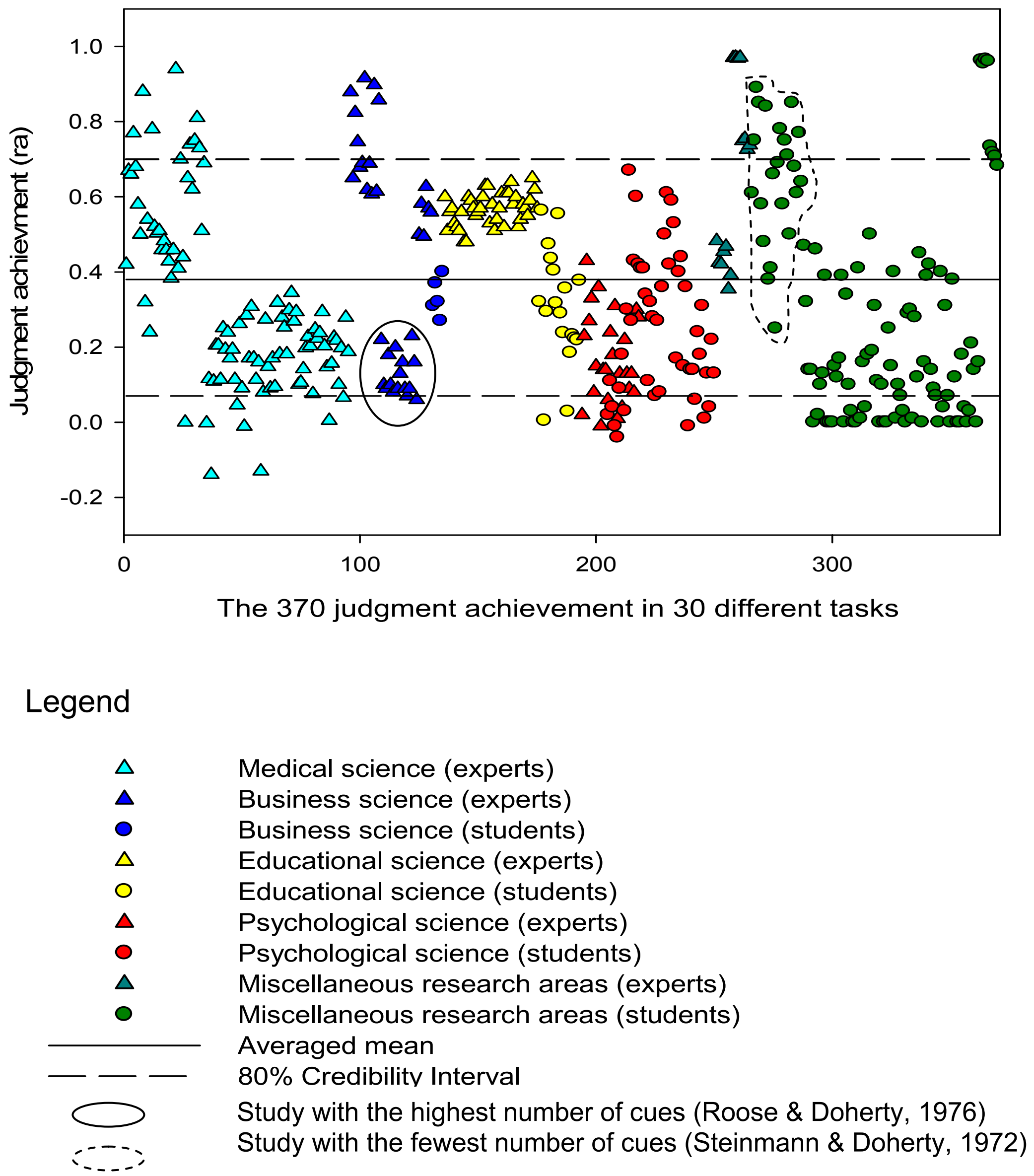
The scatter plot in Figure 9 shows clearly that the judgment achievement of individuals varies considerably. Furthermore, it shows a large 80% credibility interval for the mean from .07 to .70. The last two columns in Table 13 illustrate that the achievement correlations in our studies range from a low value of -.13 to a high level of .97. Further descriptive statistics on the overall average level of achievement correlations and on the achievement correlations separated by research areas are presented in Table 13. Looking at the second column in the last row, one can see a moderate mean of the 370 achievement correlations of .38 (see also Figure 9). But for studies applied to medical, business or psychological science, the achievement correlations are low. On the other hand, the achievement correlations increase to an almost high value in studies applied to the educational area, or to a high level in studies in other research areas. Therefore, the overall achievement correlation strongly depends on the value of the achievement correlations in studies applied to other research areas.

Research areas: As can be seen in Table 13, the achievement correlation separated by research areas is more homogenous than the overall achievement correlation, except in other research areas. By means of the scatter plots, we realized that the study by Trailer and Morgan (2004) may be responsible for the great achievement variability in studies from other research areas. Therefore, we reran the analysis and excluded this study. As expected, judgment achievement increased (r_{other} = .70; k = 45), and the variance was reduced (var_{corr} = .03), leading also to a reduction of variance in this category in comparison to the variance of .06 across studies.

Expertise within research areas: As the experience of the judges is also of interest, we checked by means of a meta-analysis. The first impression from our descriptive analysis was confirmed. There are no great differences in experts' and students' judgment achievements across areas. In addition, our analysis of expertise within research areas reveals that this tendency is not supported by educational and miscellaneous studies; in these two areas experts clearly reach better judgment achievement.

The number of used cues: In addition, Figure 9 reveals the hypothesis that the number of cues in judgment tasks can influence judgment achievement. The scatter plot shows that in the study with the highest number of cues (Roose & Doherty, 1976, see the solid outline) the subjects judged less accurately than in the study with the fewest number of cues (Steinmann & Doherty, 1972, see the dashed outline). If we consider the number of cues and exclude the study with the highest number (Roose & Doherty, 1976, see the solid outline in Figure 9), the value of the achievement correlations increases to a high value (r_a = .59), and the variation decreases (var_{corr} = .02; k = 24) in studies applied to business science.

In summary, our analysis implies that the overall achievement correlation strongly depends on the achievement values in studies applied to other research areas and to educational science.



Note. The same legend is applied to the following Figures 10 to 12.

Figure 9. The scatter plot of judgment achievement (r_a) in the 370 analyzed judgments of 30 different tasks, separated into the applied research areas. The 30 different tasks are in the same order as listed in Tables 5 and 6.

Table 13

Descriptive statistics for the separation of research areas, experience level and overall component of judgment achievement (r_a) according to a bare-bones meta-analysis (Hunter & Schmidt, 2004)

Research area:	*N*	r_a	var_{corr}	Min	Max
Medical science	95	.27	.03	-.13	.94
Business	40	.25	.04	.06	.92
Education	58	.49	.02	.01	.65
Psychology	57	.25	.00	-.04	.67
Miscellaneous	120	.52	.09	.00	.97
Experience:					
Experts[a]	196	.36	.05	-.01	.97
Business	35	.25	.05	.06	.92
Education	40	.57	.00	.48	.65
Psychology	11	.22	.00	-.01	.43
Miscellaneous	15	.73	.04	.35	.97
Students	174	.42	.07	-.04	.97
Business	5	.33	.00	.27	.40
Education	18	.30	.01	.00	.56
Psychology	46	.26	.01	-.04	.67
Miscellaneous	105	.47	.09	.00	.97
Overall	370	.38	.06	-.13	.97

Note. *N* = Corresponding to *k*, according to Hunter and Schmidt (2004, see Equation 5). r_a = weighted mean correlation according to Hunter and Schmidt (2004). var_{corr} = corrected variation according to Hunter and Schmidt (2004, variance of true score correlation). [a] includes also medical experts.

5.1.1.2 Judgment achievement components

To increase our knowledge about the underlying reason for the great heterogeneity of the reported judgment achievement values, the meta-analysis of the different *LME* components was introduced.

The G components: The scatter plot in Figure 10 reveals that the 365 analyzed judgments have a high overall average value of the component *G* (.55) as well as an increase in heterogeneity in comparison to the reported judgment achievement values (var_{corr} = .13). The average value of the component *G* in the studies separated by research area is also high, except for the low value (.29) in studies applied to business science, and the moderate value (.42) in medical science. However, the *G* component separated into different areas reduced the heterogeneity only slightly in business area, in psychology, and in other research areas (see Table 14). If we consider the experience level in our analysis, the two areas educational and other research areas – in which expert's judge higher than students – also represent high *G* components values, leading to the support of our hypothesis that high judgment achievement may be associated with high *G* component values.

The R_s component: As can be seen in Table 14, the consistency in the judgments was high (R_s = .74) in all four research areas. However, as one can see in Figure 11, the component R_s across studies also shows a substantially high variability that ranges from a low value of -.16 to a high value of .99. Finally, if we consider the R_s component in the experience level within the research areas, it is surprising that the value is only moderate for experts (R_s = .47) and high (R_s = .85) for students in psychology (see Table 16).

Like the previously reported component, *the component R_e* shows a high value across research areas. In addition, according to the pattern of the component *G*, the component R_e (.67) value is also high in studies separated by research area. If we rerun our analysis separated by the experience level within the research areas, only students in psychological science have a moderate task-predictability component, however, the increase in variability is also dominated by this subcategory.

In contrast to other components, the overall average value and also the values separated by the research area of the component *C* (.09) are quite low (see Table 14 and Appendix F: Figure 1) and without great variability in the data.

Furthermore, all components have a large 80% credibility interval (see Figures 10 - 12). If we consider the number of cues and exclude the study with the highest number (Roose & Doherty, 1976, see the solid outlines in Figures 10 - 12), all the average components are high (*G* = .77; R_s = .80; R_e = .80) in the studies applied to business science, except for one (*C* = .16), which also increased, when we considered the experience level. However, it must also be mentioned that the variation slightly increased in the consistency components.

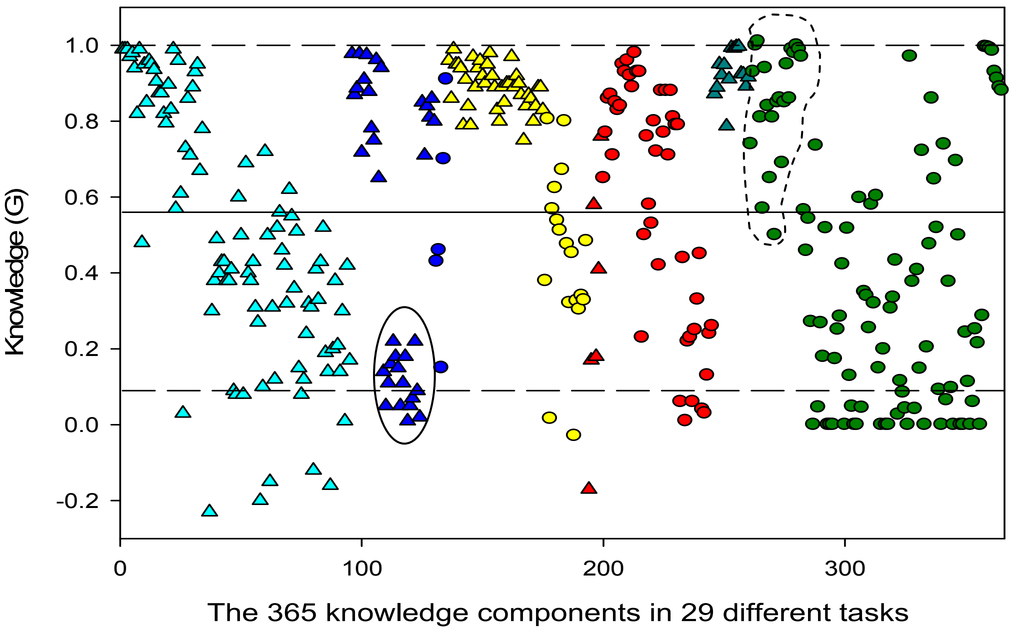
We can conclude that all underlying components of judgment achievement based on individual data also represent high heterogeneity, especially the *G* component.

Table 14

Descriptive statistics for the judgment achievement components according to a bare-bones meta-analysis (Hunter & Schmidt, 2004)

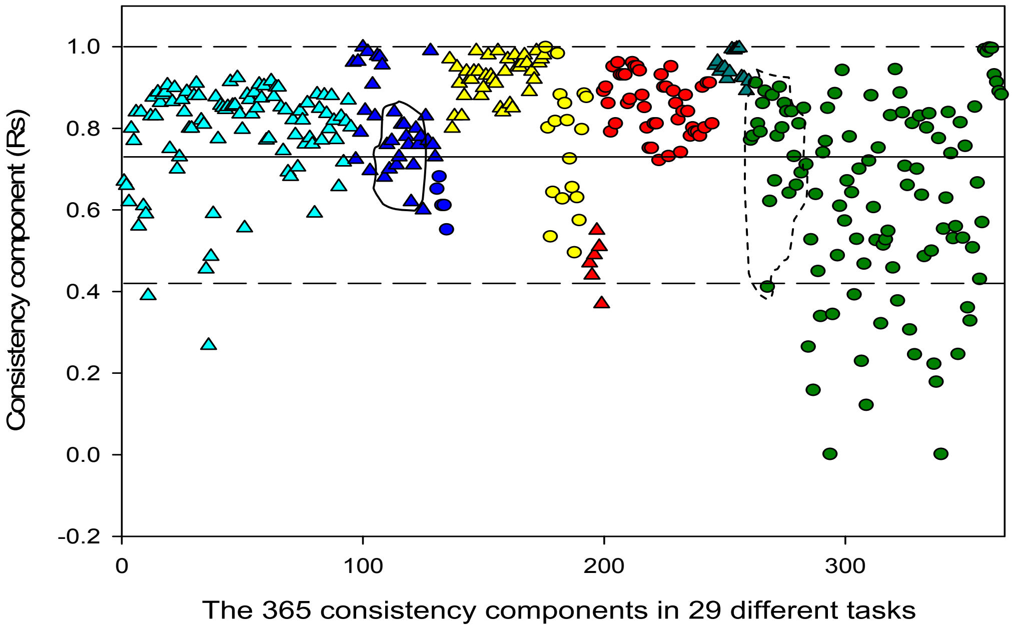
		G		R_s		R_e		C	
Research area:	N	M	var_{corr}	M	var_{corr}	M	var_{corr}	M	var_{corr}
Medical science	95	.42	.13	.79	.01	.56	.00	.10	.01
Business	40/24[a]	.29/.77[a]	.09/.03[a]	.76/.80[a]	.00/.02[a]	.53/.80[a]	.03/.02[a]	.11/.16[a]	.00/.00[a]
Education	58	.74	.06	.87	.01	.73	.00	.01	.00
Psychology	57/52[b]	.55[b]	.08[b]	.82[b]	.01[b]	.53	.08	.06[b]	.01[b]
Miscellaneous	120	.68	.11	.56	.11	.82	.02	.12	.03
Experience:									
Experts	196[b, c]	.53[b]	.14[b]	.83[b]	.01[b]	.60	.02	.10[b]	.01[b]
Business	35/19[a]	.27/.85[a]	.08/.00[a]	.77/.85[a]	.00/.01[a]	.53/.88[a]	.03/.01[a]	.10/.14[a]	.00/.00[a]
Education	40	.88	.00	.93	.00	.69	.00	.00	.00
Psychology	11/6[b]	.39[b]	.10[b]	.47[b]	.00[b]	.72	.00	.25[b]	.00[b]
Miscellaneous	15	.94	.00	.96	.00	.77	.04	.27	.02
Students	174	.59	.11	.57	.08	.78	.04	.08	.02
Business	5	.53	.05	.62	.00	.59	.00	.23	.00
Education	18	.43	.04	.75	.02	.84	.00	.03	.00
Psychology	46	.57	.07	.85	.00	.49	.08	.03	.01
Miscellaneous	105	.63	.12	.48	.09	.84	.02	.10	.03
Overall	370/365[b]	.55[b]	.13[b]	.74[b]	.06[b]	.67	.03	.09[b]	.01[b]

Note. N = Corresponding to k, according to Hunter and Schmidt (2004, see Equation 5). var_{corr} = corrected variation according to Hunter and Schmidt (2004, variance of true score correlation). [a]rerun meta-analysis with the exclusion of the study by Roose and Doherty (1976). [b]the difference in N is based on the study by Werner et al. (1989), as in this study the consistency and the knowledge components was not available at the individual level, which leads to 5 missing in thee components. [c]includes also medical expert.



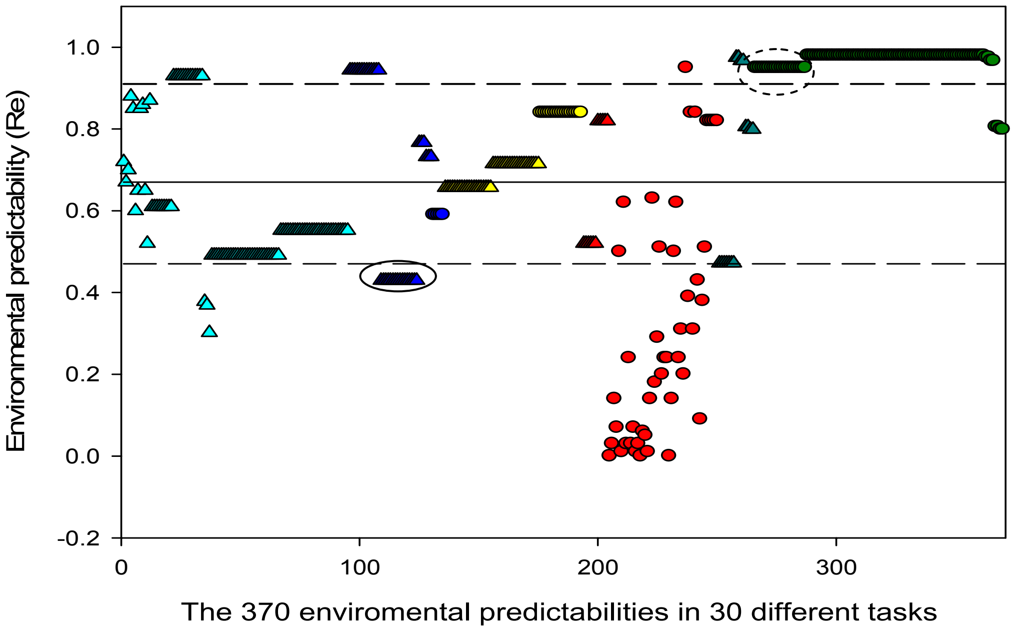
You will find the legend on page 99.

Figure 10. The scatter plot of the knowledge component (*G*) in the 365 analyzed judgments in 29 different tasks, separated into the applied research areas. The 29 different tasks are in the same order as listed in Tables 5 and 6.



You will find the legend on page 99.

Figure 11. The scatter plot of the consistency component (R_s) in the 365 analyzed judgments in 29 different tasks, separated into the applied research areas. The 29 different tasks are in the same order as listed in Tables 5 and 6.



You will find the legend on page 99.

Figure 12. The scatter plot of the environmental predictability component (R_e) in the 370 analyzed judgments in 30 different tasks, separated into the applied research areas. The 30 different tasks are in the same order as listed in Tables 5 and 6.

5.1.2 Psychometric meta-analysis

In the following, our results of a psychometric meta-analysis based on individual data are described. For an overview of our correction we refer to chapter 4.5.

5.1.2.1 Judgment achievement

In the following Table 15, the psychometric meta-analysis based on individual data is presented. As noted previously, in educational, psychological and the miscellaneous area, there were no data or retest-reliability values available for our measurement-error correction; hence, as we assume that in every study a measurement error is included, we made three different estimations: We assumed a retest-reliability value of .78 (see Ashton, 2000), and two extreme retest-reliability values of .90 and .50 are used for our measurement-error calculation.

Table 15 presents the average judgment achievement corrected for measurement error. Judgment achievement across different research areas increased from a moderate value .38 to a minimum level of .46, and, finally, to a high level of .65. However, the variability pattern found in our previous bare-bones meta-analysis remains.

Expertise. Also in the psychometric meta-analysis, our hypothesis that experts judge better than non-experts across all research areas, although their judgment is measurement error corrected, is not confirmed. However, a closer look at the data reveals that there are again domain differences supporting the hypothesis of differences between research areas (see Tables 16, 17).

In summary, our results found with a bare-bones meta-analysis are confirmed. In addition, this analysis also shows that a simple bare-bones meta-analysis clearly underestimates judgment achievement. However, to shed light on the underlying reasons of judgment accuracy or inaccuracy we present a psychometric meta-analysis of the remaining *LME* components in the following.

5.1.2.2 Judgment achievement components

The *G* component shows an increase from .55 to minimal .67 to .94 across different research areas. Hence, in a psychometric meta-analysis the *G* component increased with a minimum of 12%. However, if we look at the different research areas, our analysis reveals differences, especially the knowledge component in medical science increases from a moderate level (.42) in a bare-bones meta-analysis to a high level (.57) in a psychometric meta-analysis (see Table 15). Furthermore, the experience level again represents the previous bare-bones meta-analysis, however, the level increased clearly, as both experts and non-experts knowledge components increased (see Tables 16, 17).

The consistency component. In a psychometric meta-analysis the consistency component increased with a minimum of 5% (R_s = .79) in the .90 retest-reliability correction to 19% (R_s = .95) if we assume .50 retest-reliability. However, if we look at the differences between research areas, there is only a slight increase to be found in other research areas 3% (R_s = .59) at the minimum assuming a .90 retest-reliability across all research areas (see Table 15). Finally, also experts in psychology science reach a high consistency level (R_s = .50) if we assume a conservative .90 retest-reliability value for our measurement corrections. However, as can be clearly noticed, there is almost no variation in experts' consistency components within the different research areas. On the other hand, the variation is dominant in student consistency in other research areas (see Tables 16, 17).

The environmental predictability components. Our psychometric meta-analysis reveals high task predictability conditions across areas as well as within research areas (see Table 15). Furthermore, there is also no difference between experts and student tasks. Both also reach a high value between the different research areas. Hence, student tasks in psychological science increased from a moderate value (R_e = .49) in a bare-bones meta-analysis to a high value (R_e = .52) in a psychometric meta-analysis. However, the great variations in this category remain

(var_{corr} = .10) and dominate the overall variation across research areas (var_{corr} = .05, minimal, see Tables 16, 17).

The non-linear knowledge components. In comparison to the presented components the *C* component has the smallest increase in a psychometric meta-analysis, or remains stable in a correction with a retest-reliability value of .90 (see Table 15). However, the slight differences between experts' and students' non-linear knowledge components imply that experts have slightly higher values across areas, and clearly higher values in psychological and other research areas. It must also be mentioned that experts in business science (*C* = .10) reach a lower level than business science students (*C* = .25), but both still have low non-linear knowledge components (see Tables 16, 17).

Summing up our psychometric meta-analysis on the *LME* components based on individual data, we conclude that all values increased, but the heterogeneity still remains.

Table 15

Descriptive statistics for the separation of research areas and overall components of correlations of the LME according to a psychometric meta-analysis (Hunter & Schmidt, 2004)

			r_a		G		R_s		R_e		C	
Research area:	*rr*	*N*	*M*	var_{corr}	*M*	var_{corr}	*M*	var_{corr}	*M*	var_{corr}	*M*	var_{corr}
Medical science		95	.36	.05	.57	.24	.91	.05	.66	.01	.13	.01
Business		40	.28	.05	.31	.11	.84	.00	^{b}	^{b}	.11	.00
Education	.90	58	.54	.02	.83	.07	.92	.02	.77	.00	.01	.00
	.78		.62	.03	.96	.09	.98	.02	.83	.00	.02	.00
	.50		.97	.07	--	--	--	--	--	--	.03	.00
Psychology	.90	57^a	.28	.01	$.62^a$	$.10^a$	$.87^a$	$.01^a$	.55	.09	$.05^a$	$.01^a$
	.78		.32	.01	$.71^a$	$.14^a$	$.93^a$	$.01^a$	.60	.10	$.06^a$	$.01^a$
	.50		.50	.03	--	--	--	--	.75	.16	$.08^a$	$.02^a$
Miscellaneous	.90	120	.58	.11	.75	.14	.59	.13	.87	.03	.10	.03
	.78		.66	.15	.87	.19	.63	.14	.93	.03	.12	.02
	.50		--		--	--	.79	.22	--	--	.18	.09
Overall	.90	370^a	.46	.09	$.67^a$	$.18^a$	$.79^a$	$.06^a$	.74	.04	$.11^a$	$.02^a$
	.78		.50	.10	$.73^a$	$.22^a$	$.83^a$	$.07^a$	.77	.05	$.11^a$	$.02^a$
	.50		.65	.16	$.93^a$	$.35^a$	$.95^a$	$.09^a$	.87	.05	$.13^a$	$.03^a$

Note. rr = Suggested retest-reliability values for our measurement error corrections. *N* = Corresponding to *k*, according to Hunter and Schmidt (2004, see Equation 5). var_{corr} = corrected variation according to Hunter and Schmidt (2004, variance of true score *correlation).* -- = value greater than 1. [a]In the study by Werner et al. (1989) the consistency and the knowledge component were not available at the individual level, which leads to 5 missing values in these components. [b]see bare-bones meta-analysis, no correction because this category includes only objective criterions.

Table 16

Descriptive statistics for experts in relation to the separation of research areas and overall components of correlations of the LME according to a psychometric meta-analysis (Hunter & Schmidt, 2004)

			r_a		G		R_s		R_e		C	
Research area:	*rr*	*N*	*M*	var_{corr}	*M*	var_{corr}	*M*	var_{corr}	*M*	var_{corr}	*M*	var_{corr}
Business		35	.27	.06	.30	.11	.85	.00	[b]	[b]	.10	.00
Education	.90	40	.63	.00	.99	.00	.98	.00	.72	.00	.00	.00
	.78		.73	.00	--	--	--	--	.83	.00	.00	.00
	.50		--	--	--	--	--	--	.97	.00	.01	.00
Psychology	.90	11[a]	.23	.00	.40[a]	.11[a]	.60[a]	.00[a]	[b]	[b]	.26[a]	.00[a]
	.78		.25	.00	.43[a]	.13[a]	.64[a]	.00[a]	[b]	[b]	.28[a]	.00[a]
	.50		.31	.00	.55[a]	.20[a]	.80[a]	.00[a]	[b]	[b]	.35[a]	.00[a]
Miscellaneous	.93[c]	15	.78	.05	.99	.00	--	--	[b]	[b]	.29	.02
Overall	.90	196[a,d]	.46	.08	.68[a]	.23[a]	.92[a]	.01[a]	.69	.02	.12[a]	.01[a]
	.78		.48	.09	.72[a]	.25[a]	.94[a]	.01[a]	.70	.02	.13[a]	.01[a]
	.50		.55	.11	.81[a]	.31[a]	1.00[a]	.01[a]	.75	.02	.14[a]	.02[a]

Note. rr = Suggested retest-reliability values for our measurement error corrections. *N* = Corresponding to *k*, according to Hunter and Schmidt (2004, see Equation 5). var_{corr} = corrected variation according to Hunter and Schmidt (2004, variance of true score correlation). -- = value greater than 1. [a]In the study by Werner et al. (1989) the consistency and the knowledge component were not available at the individual level, which leads to 5 missing values in these components. [b]see bare-bones meta-analysis, no correction because this category includes only objective criterions. [c]no further correction because only meteorologists are included (*rr* = .93, Ashton, 2000). [d]includes also medical experts.

Table 17

Descriptive statistics for students in relation to the separation of research areas and overall components of correlations of the LME according a psychometric meta-analysis (Hunter & Schmidt, 2004)

			r_a		G		R_s		R_e		C	
Research area:	*rr*	*N*	*M*	var_{corr}	*M*	var_{corr}	*M*	var_{corr}	*M*	var_{corr}	*M*	var_{corr}
Business		5	.36	.00	--	--	.69	.00	[a]	[a]	.25	.00
Education	.90	18	.34	.02	.49	.05	.80	.02	.89	.00	.04	.00
	.78		.39	.02	.56	.07	.85	.02	.95	.00	.04	.00
	.50		.61	.06	.88	.16	--	--	--	--	.06	.01
Psychology	.90	46	.28	.01	.63	.09	.89	.00	.52	.10	.04	.01
	.78		.33	.02	.73	.13	.96	.00	.55	.11	.04	.01
	.50		.52	.04	--	--	--	--	.69	.17	.07	.02
Miscellaneous	.90	105	.52	.11	.70	.14	.51	.11	.88	.02	.08	.03
	.78		.61	.15	.80	.20	.54	.12	--	--	.13	.05
	.50		.94	.36	--	--	.68	.19	--	--	.15	.09
Overall	.90	174	.46	.09	.65	.13	.61	.10	.83	.04	.07	.02
	.78		.53	.12	.75	.17	.65	.12	.89	.05	.08	.03
	.50		.82	.28	--	--	.81	.18	--	--	.13	.07

Note. rr = Suggested retest-reliability values for our measurement error corrections. *N* = Corresponding to *k*, according to Hunter and Schmidt (2004, see Equation 5). var_{corr} = corrected variation according to Hunter and Schmidt (2004, variance of true score correlation). -- = value greater than 1. [a]see bare-bones meta-analysis, no correction because this category includes only objective criterions.

5.1.3 Intercorrelations of the components

To enhance our knowledge about the underlying reasons in judgment achievement, we also considered its intercorrelations. The intercorrelation across research areas (see Table 18) and within research areas (see Table 19) is presented. At first glance, judgment achievement significantly correlates with every component (see Table 18). There is, however, a negative correlation between the knowledge and the environment component (-.02), which implies that task predictability is negatively associated with knowledge. However, if we separate our data base into experience levels (see Appendix F: Tables 2, 3), our results reveal that the negative correlation between knowledge and task validity remains in the student data base – and increases to a high level in experts' judgment achievement, except when it comes to educational experts (-.44). However, as becomes obvious, there are a lot of missing values due to small sample size. Hence, the reported intercorrelation should be interpreted with caution (see Appendix F: Tables 2, 3).

Table 18

Intercorrelation of the LME components

Components	Components				
Overall	r_a	*G*	R_s	R_e	*C*
r_a	--	.84**	.50**	.25**	.38**
G	.84**	--	.47**	-.02	.10
R_s	.50**	.47**	--	-.27**	.06
R_e	.25**	-.02	-.27**	--	.09
C	.38**	.10	.06	.09	--
Experts					
r_a	--	.87**	.46**	.79**	.27**
G	.87**	--	.47**	.65**	.01
R_s	.46**	.47**	--	.34**	-.15*
R_e	.79**	.65**	.34**	--	.21**
C	.27**	.01	-.15*	.21**	--
Students					
r_a	--	.79**	.49**	.07	.45**
G	.79**	--	.45**	-.40**	.14
R_e	.49**	.45**	--	-.40**	.10
R_s	.07	-.40**	-.40**	--	.17*
C	.45**	.14	.10	17*	--

Note. ** Correlation is significant at the .001 level (2-tailes).

* Correlation is significant at the .005 level (2-tailes).

Table 19

Intercorrelation of the LME components in the different areas

Components in:	Components				
Medical science	r_a	*G*	R_s	R_e	*C*
r_a	--	.85**	.14	.79**	.47**
G	.85**	--	.22*	.60**	.16
R_s	.14	.22*	--	.14	-.08
R_e	.79**	.60**	.14	--	.31**
C	.47**	.16	-.08	.31**	--
Business science					
r_a	--	.93**	.60**	.96**	.07
G	.93**	--	.37*	.91**	.01
R_s	.60**	.37*	--	.54**	-.24
R_e	.96**	.91**	.54**	--	.05
C	.07	.01	-.24	.05	--
Education science					
r_a	--	.96**	.80**	-.74**	-.07
G	.96**	--	.70**	-.83**	-.18
R_s	.80**	.70**	--	-.60**	-.30*
R_e	-.74**	-.83**	-.60**	--	.10
C	-.07	-.18	-.30*	.10	
Psychology science					
r_a	--	.44**	.14	.12	.28*
G	.44**	--	.40**	-.62**	-.35*
R_s	.14	.40**	--	-.26	-.43**
R_e	.12	-.62**	-.26	--	.42**
C	.28*	-.35*	-.43**	.42**	--
Miscellaneous					
r_a	--	.92**	.68**	-.23*	.69**
G	.92**	--	.55**	-.42**	.54**
R_e	.68**	.55**	--	-.39**	.44**
R_s	-.23*	-.42**	-.39**	--	-.17
C	.69**	.54**	.44**	-.17	--

Note. ** Correlation is significant at the .001 level (2-tailes).

* Correlation is significant at the .005 level (2-tailes).

In summary, our results based on *LME* components for individuals lead to a small sample. Therefore, our results must be accepted with caution. Hence, we will supplement our data with studies including *LME* components across individuals (or nomothetic data bases) in the following.

5.2 Nomothetic data base

The introduced meta-analysis based on individual data is supplemented by studies including only nomothetic data. In line with the previous meta-analysis, we will first present our results with a bare-bones meta-analysis and then with a psychometric meta-analysis.

5.2.1 Bare-bones meta-analysis

The following meta-analytic results are presented in two sections. The first section describes the results for the achievement correlations across the judgment tasks presented in Table 20 and Figure 13. The second section reveals the additional correlations of components of the *LME* across the judgment tasks in Tables 21 to 23 and Figures 14 to 17.

5.2.1.1 Judgment achievement

The achievement correlations are summarized in Table 20 and Figure 13. There was a moderate mean (.40) from the 49 achievement correlations across 1151 analyzed judgment achievements by 1055 judges. The 75% rule indicates that there were true differences in effect sizes across judgment tasks. Accordingly, separate meta-analyses were calculated for categories of studies like the research area and the experience level in the different research areas.

Table 20

Bare-bones meta-analysis according to the method of Hunter-Schmidt (2004) supplemented by a trim-and-fill analysis on judgment achievement (r_a), separated into research area and experience level

Research area	*k*	*N*	r_a	var_{corr}	80% CI		75%
Medicine	10/11	258/262	.40/.39	.00/.oo	.40/.38	.40/.38	157/134
Business	9/13	239/332	.50/.19	.07/.25	.15/-.45	.84/.84	24.45/13.56
Education	4/5	156/176	.39/.37	.00/.02	.39/.13	.39/.50	177.89/53.75
Psychology	14/15	249/257	.22/.21	.00/.00	.22/.20	.22/.20	448.50/319
Miscellaneous	12/17	249/291	.44/.37	.02/.07	.28/.00	.61/.66	67.55/43.98
Overall	49/58	1151/1285	.39/.30	.02/.07	.23/-.04	.55/.64	69.42/36.49
Experts in:							
Business	6/9	116/136	.36/.25	.01/.05	.26/-.03	.46/.52	87.73/60.24
Education	2	40	.57	.00	.57	.57	975.69
Psychology	4/6	59/70	.10/.06	.00/.00	.10/.06	.10/.06	975.77/635.55
Miscellaneous	5/7	15/23	.65/.30	.00/.00	.65/.30	.65/.30	401.60/158.46
Overall[a]	27/32	488/518	.37/.32	.00/.01	.37/.19	.37/.46	129.00/84.6
Students in:							
Business	3	123	.63	.10	.23	1.00	8.52
Education	2	116	.33	.00	.33	.33	27143
Psychology	10	190	.26	.00	.26	.26	606
Miscellaneous	11/16	234/279	.43/.31	.01/.06	.33/.00	.53/.00	86.40/52.59
Overall	25/29	663/695	.40/.41	.02/.41	.21/.41	.59/.76	58.94/40.28

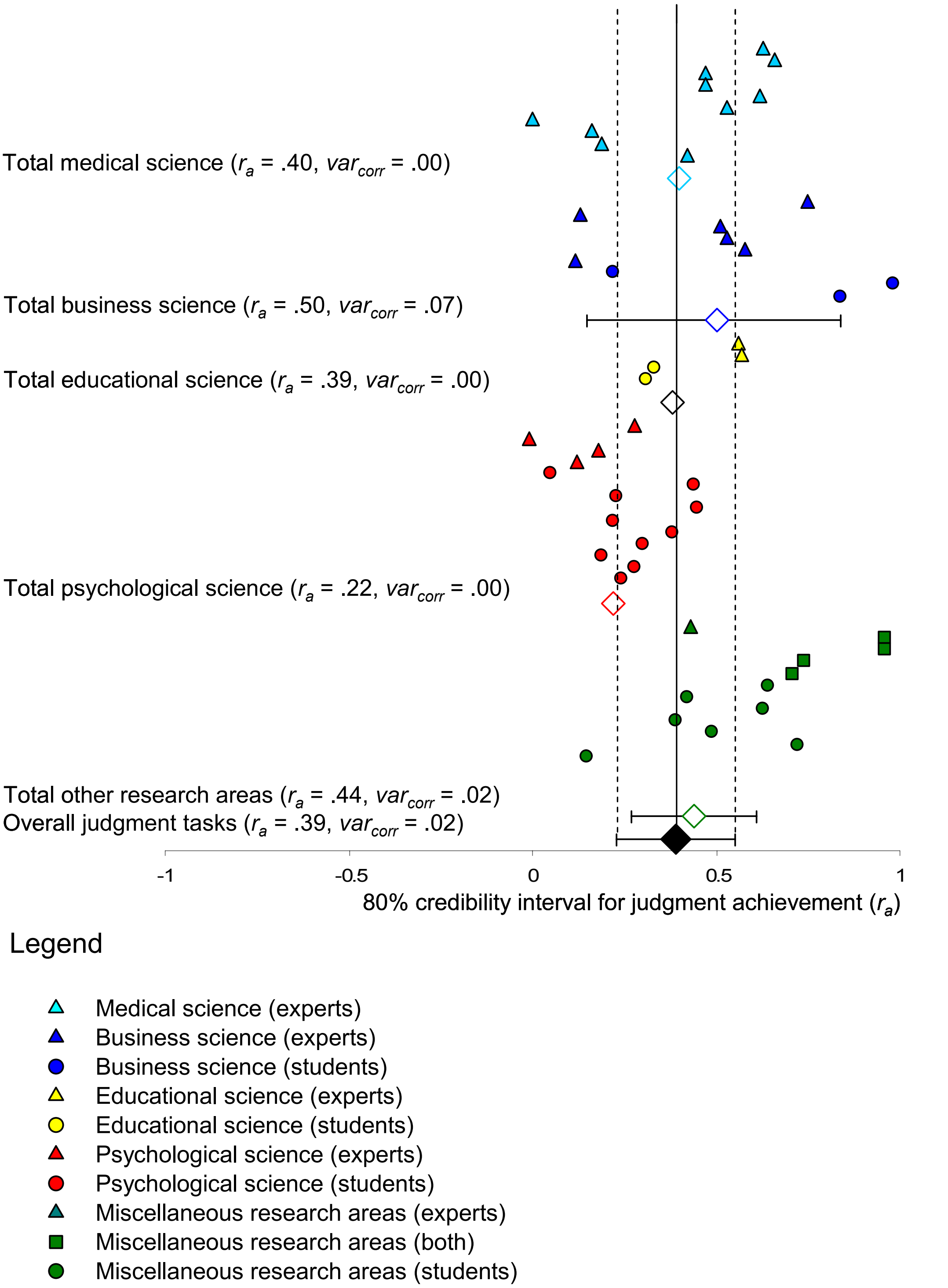
Note. *k* = number of correlations (i.e. judgment tasks). *N* = total sample size for all judgment tasks combined. r_a = weighted mean correlation according to Hunter and Schmidt (2004). var_{corr} = corrected variation according to Hunter and Schmidt (2004, variance of true score correlation). 80% CI = 80% credibility interval for true score correlation distribution. 75% = Percentage variance of observed correlations due to all artefacts, if below 75%, it indicates moderator variable. /Fill-and-trim analysis results after a publication bias is indicated. [a]this analysis includes medical experts. Grey boxes: Results not confirmed by the trim-and-fill analysis.

The achievement correlations were lowest in psychology (r_a = .22) and increased for studies applied to the educational (r_a = .39), the medicine (r_a = .40) and the miscellaneous professional area (r_a = .44), and to a higher value for studies in business areas (r_a = .50), resulting in the highest level of achievement. In addition, the 75% rule indicates moderating variables not only across studies, but also in the meta-analyses of the sub-group of business and other research area studies, or the two research areas with the highest judgment achievements.

Furthermore, it is clear that the greatest variability is found in business-sciences judgment achievement. We reran the analysis, however, separating the experience level of the judges. This separation revealed that in experts' judgment achievement across or within research areas no moderator variables were indicated. On the other hand, it is also clear, that students' judgment achievement in business sciences is responsible for the moderator variable indication in students' judgment achievement across research areas.

Finally, our trim-and-fill application when a publication bias was indicated confirms our results with some exceptions, such as in business science. In this category, the suggested judgment achievement values decreased from a high value of .50 to a low value of .19. This is explained by experts' judgment achievement, as there was no publication bias indicated in studies using business students. In the same way, there is a decrease in experts' judgment achievement in other research areas to a moderate level. Although the judgment-achievement values for students in other research areas is stable, there are now moderator variables indicated. It must also be mentioned that after a publication-bias correction judgment achievement in educational science indicated moderator variables, but it despairs after we separated the analysis according to the experience level.

In the following, the additional components are considered, in order to clarify the underlying reasons for the reported achievement values.



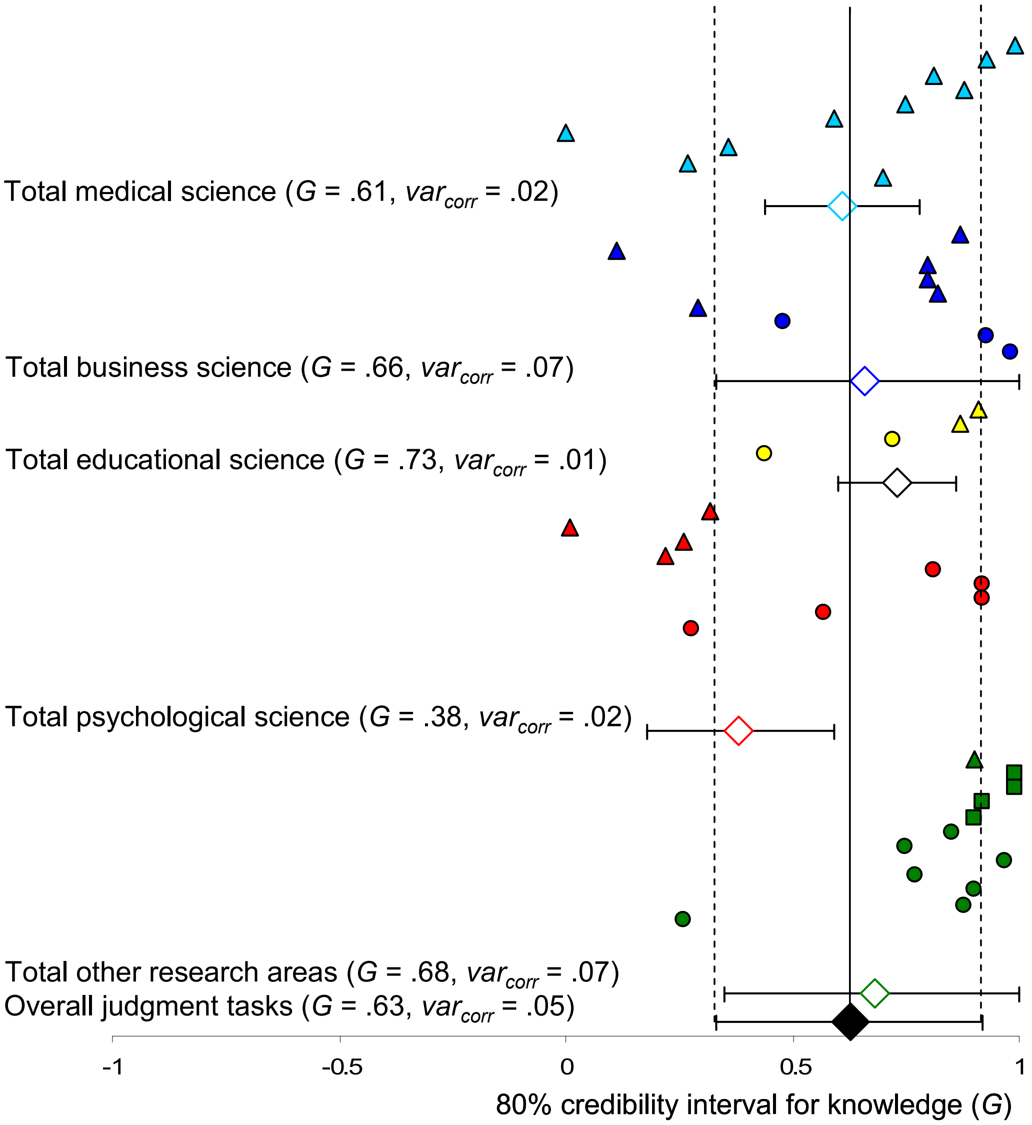
Note. The same legend is applied to the following Figures 14 to 17.

Figure 13. The forest plot of judgment achievement (r_a), separated into the applied research areas, and within these into experience levels. The studies in the forest plots are in the same order as in Table 5 and 6.

5.2.1.2 Judgment achievement components

Knowledge components. A high average value of the knowledge component *G* (.63) is presented in Figure 14 and Table 21 (note that the sample sizes vary, because some components of the *LME* could not be calculated). The overall meta-analysis of the *G* component indicates moderator variables. However, also a separation into research areas indicates moderator variables in each area. The average value of the component *G* in the studies separated by the research areas is high, except for the moderate value (.38) in psychological studies. Hence, we reran the analysis, separating the experience level within research areas. Against our expectation, students' knowledge across research areas is higher than experts' knowledge, but both components are high and also indicate moderator variables. A look at the different research areas, however, shows that the knowledge component decreases from a moderate value (.38) across areas to a low level (.17) in psychological experts' knowledge component, leading to the only value which is not high. Finally, the moderator-variable indication in the experts' knowledge component is dominated by business and medical sciences experts. An inspection of the scatter plot of correlations suggested the exclusion of the study by Roose and Doherty (1976) and Mear and Firth (1987), both studies are also the only ones with low judgment achievement in the business category, which leads to the hypothesis that maybe judgment achievement is associated with a low knowledge component. After the exclusion of these two studies, no moderator factors were evident (G = .81; var_{corr} = .00; k = 4; N = 62). The exclusion of the study by Roose and Doherty (1976) with a large number of 64 cues also supports the view that the extreme number of cues enhanced the variability of the data.

Finally, our trim-and-fill method application if a publication bias is indicated reveals that the knowledge analysis is robust against it, except in the category of psychology students and other research area, leading to a moderate instead of a high knowledge component.



You will find the legend on page 119.

Figure 14. The forest plot of the knowledge component (G), separated into the applied research areas, and within these by the experience level. The studies in the forest plots are in the same order as in Table 5 and 6.

Table 21

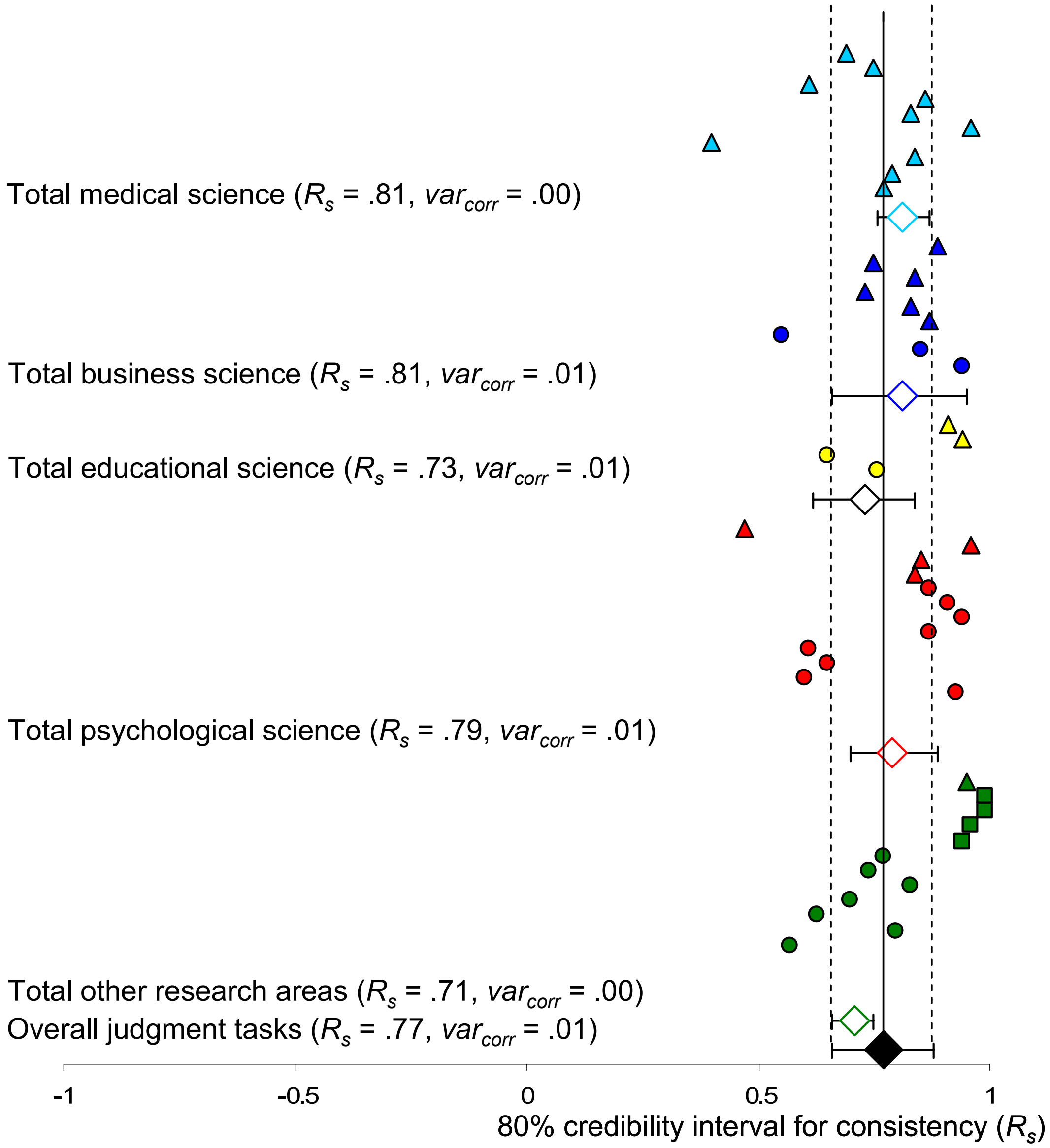
Bare-bones meta-analysis according to the method of Hunter-Schmidt (2004), supplemented by a trim-and-fill analysis of the knowledge component (G), separated into research areas and experience levels

Research area	*k*	*N*	*G*	var_{corr}	80% CI		75%
Medicine	10/11	258/262	.61/.59	.02/.02	.44/.37	.78/.80	50.72/41.85
Business	9	239	.66	.07	.33	1.00	15.85
Education	4	156	.73	.01	.60	.86	35.62
Psychology	9/11	105/121	.38/.24	.02/.11	.17/-.17	.58/.66	73.77/45.46
Miscellaneous	12/17	249/313	.68/.49	.07/.16	35/-.02	1.00/1.0	19.26/17.90
Overall	44/47	1007/1019	.63/.63	.05/.06	.34/.30	.93/.93	24.91/23.34
Experts in:							
Business	6/7	116/129	.55/.45	.05/.11	.25/.02	.84/.87	33.56/26.45
Education	2	40	.89	.00	.89	.89	313.80
Psychology	4/5	59/65	.17/.13	.00/.00	.17/.14	.17/.14	444.93/302
Miscellaneous	5/6	15/19	.92/.71	.00/.03	.92/.48	.92/.94	768.55/80.9
Overall[a]	27/32	488/508	.57/.53	.04/.06	.32/.20	.83/.85	43.69/38.55
Students in:							
Business	3	123	.78	.05	.49	1.00	6.95
Education	2	116	.68	.00	.59	.77	51.03
Psychology	5/7	46/62	.65/.37	.03/.17	.42/-.16	.87/.90	57.65/35.14
Miscellaneous	11/15	234/271	.66/.53	.06/.13	.34/.07	.98/1.00	24.15/22.03
Overall	21/27	519/631	.69/.52	.04/.16	.41/.00	.97/1.00	21.81/14.41

Note. *k* = number of correlations (i.e. judgment tasks). *N* = total sample size for all judgment tasks combined. *G* = weighted mean correlation according to Hunter and Schmidt (2004). var_{corr} = corrected variation according to Hunter and Schmidt (2004, variance of true score correlation). 80% CI = 80% credibility interval for true score correlation distribution. 75% = Percentage variance of observed correlations due to all artefacts, if below 75%, it indicates moderator variable. /Results of the fill-and-trim analysis after a publication bias is indicated. [a]this analysis includes medical experts. Grey boxes: Results not confirmed by the trim-and-fill analysis.

Consistency. Figure 15 and Table 22 indicate that on average the subjects were highly consistent in their judgments (R_s = .77). The 75% rule indicates a lack of homogeneity of the single effect sizes, and further meta-analyses were conducted. Moderator factors are indicated for studies related to all research areas, except for studies in other research areas. Hence, we reran the analysis, separating the experience level within research areas. Although the overall expert-consistency component indicated no moderator variables, psychology and medical experts' consistency indicated moderator variables. A scatter plot of medical experts' consistency component, however, reveals a low value of the three physicians in the study by Einhorn (1974). In a following meta-analysis of medical experts, with the exclusion of Einhorns study, no moderator variables are evident (R_s = .81; var_{corr} = .00; k = 9; N = 255). Although scatter plots of experts in business science were created, no possible judgment tasks could be identified, as all values are high. Finally, across research areas, students' consistency is clearly dominated by students in business sciences. However, scatter plots of the three included judgment tasks indicate that all values are high, and thus, no judgment task could be identified for a possible exclusion in a reanalysis.

Finally, the moderator variable indicated in our publication-bias analysis supplemented by the fill-and-trim method reveals no influence in the consistency component, as all consistency values are still high. However, this analysis leads to moderator indications in experts' consistency component based mainly on the values of experts' consistency in the other research areas. In addition, there are moderator variables indicated in the psychology student's category.



You will find the legend on page 119.

Figure 15. The forest plot of the consistency component (R_s), separated into the applied research areas, and within these by the experience level. The studies in the forest plots are in the same order as in Table 5 and 6.

Table 22

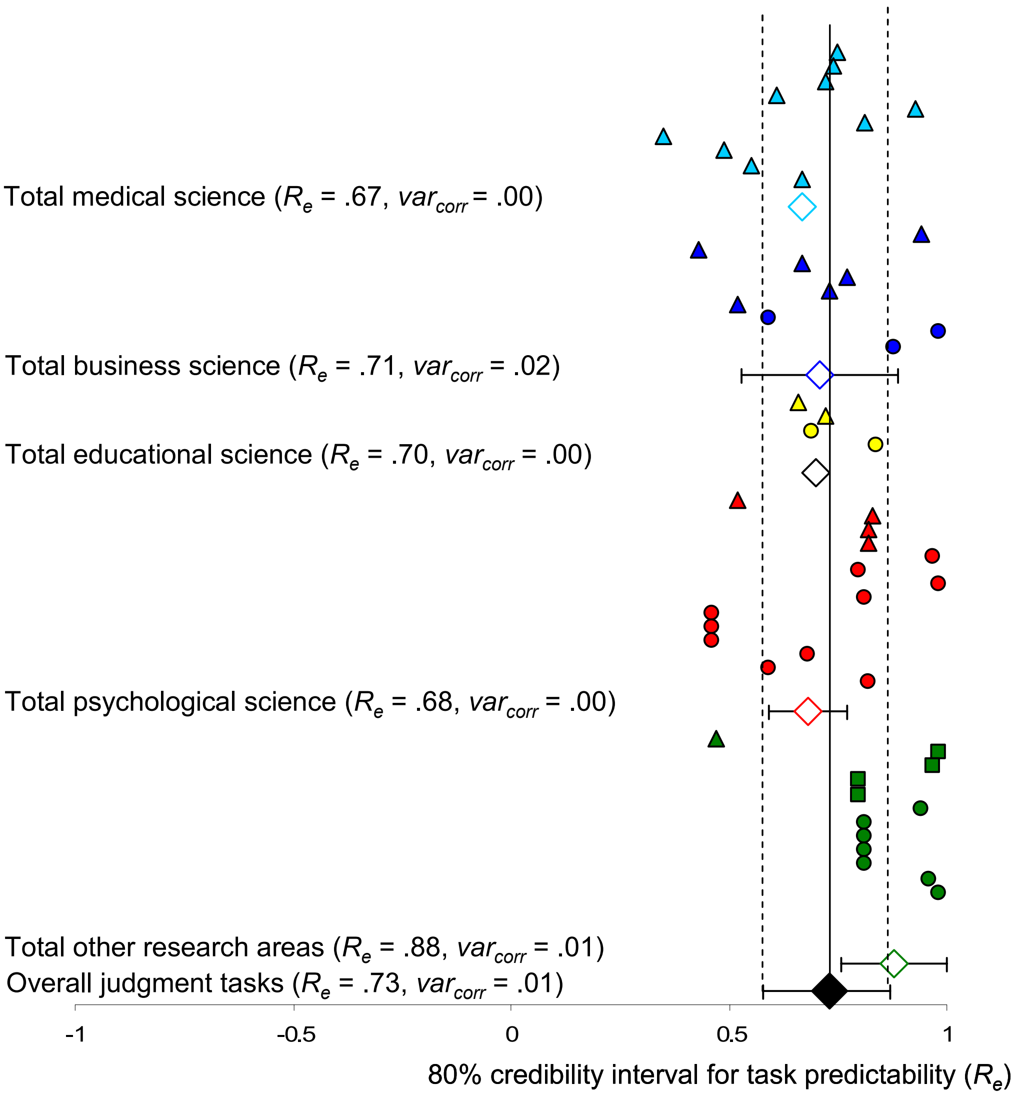
Bare-bones meta-analysis according to the method of Hunter-Schmidt (2004), supplemented by a trim-and-fill analysis of the consistency component (R_s), separated into research areas and experience levels

Research area	*k*	*N*	R_s	var_{corr}	80% CI		75%
Medicine	10/12	258/265	.81/.79	.00/.00	.75/.68	.86/.89	74.95/53.63
Business	9/11	239/303	.81/.67	.01/.06	.66/.33	.95/1.00	28.60/15.00
Education	4/6	156/196	.73/.53	.01/.14	.62/.04	.84/1.00	43.52/9.90
Psychology	12	150	.79	.01	.69	.88	71.34
Miscellaneous	12/17	249/272	.71/.64	.00/.05	.66/.34	.75/.92	90.86/34.01
Overall	47/58	1052/1260	.77/.61	.01/.12	.66/.16	.88/1.05	53.34/14.61
Experts in:							
Business	6/7	116/119	.84/.61	.00/.00	.84/.60	.84/.60	268.23/105.37
Education	2	40	.92	.00	.92	.92	1241.73
Psychology	4/5	59/65	.85/.80	.01/.02	.75/.60	.95/.99	48.83/33.04
Miscellaneous	5/6	15/19	.95/.75	.00/.05	.95/.45	.95/1.00	1724.68/66.77
Overall[a]	27/29	488/496	.83/.81	.00/.01	.80/.66	.87/.94	89.61/37.21
Students in:							
Business	3	123	.77	.03	.56	.98	12.68
Education	2	116	.66	.00	.66	.66	422.27
Psychology	8/11	91/115	.74/.57	.00/.09	.74/.18	.74/.95	107.28/35.15
Miscellaneous	11	234	.69/.63	.00/.00	.69/.52	.69/.73	148.50/80.0
Overall	17/33	399/664	.70/.56	.01/.10	.60/.15	.80/.97	69.27/20.84

Note. k = number of correlations (i.e. judgment tasks). *N* = total sample size for all judgment tasks combined. R_s = weighted mean correlation according to Hunter and Schmidt (2004). var_{corr} = corrected variation according to Hunter and Schmidt (2004, variance of true score correlation). 80% CI = 80% credibility interval for true score correlation distribution. 75% = Percentage variance of observed correlations due to all artefacts, if below 75%, it indicates moderator variable. /Results of the fill-and-trim analyses after a publication bias is indicated. [a]this analysis includes medical experts. Grey boxes: Results not confirmed by the trim-and-fill analysis.

Environmental predictability. The overall level of the environmental predictability component R_e (.73) was high (see Figure 16 and Table 23). The 75% rule also indicates the presence of moderated relationships in the environmental-predictability component. Further analyses separating correlations into research areas were conducted. The largest relationship was found between environmental predictability and studies from the miscellaneous research area (R_e = .88). The largest variation of component R_e is in business studies, but this area has the largest range of cues (up to 64 cues) of all the categories. However, again, all task predictability values are high, implying no research-area differences in the type of task. On the other hand, the 75% rule indicates moderator variables for the studies from the business or the miscellaneous research area. An additional meta-analysis under exclusion of studies could not identify judgment tasks with possible moderator variables in this category. Hence, we reran our analysis, separating the experience level in studies within research areas. Although experts' task predictability is lower than students' task predictability, they are both still high. Furthermore, experts' task predictability indicated no moderator variables in comparison to students' task predictability. A closer look at the scatter plots of students' task predictability in business and other research areas, which indicated moderator variables, reveals that all included values are high. Thus, we could not identify any task characteristics which could influence our results.

Finally, after a trim-and-fill application if a publication bias is indicated in the psychology category, moderator values are revealed which can't be explained by the experience level. In addition, the high value in experts' task predictability in other research areas reaches a moderate value. Finally, although the business experts' task-predictability component is stable, there are now moderator variable indicated.



You will find the legend on page 119.

Figure 16. The forest plot of the task-predictability component (R_e), separated into the applied research areas, and within these by experience level. The studies in the forest plots are in the same order as in Table 5 and 6.

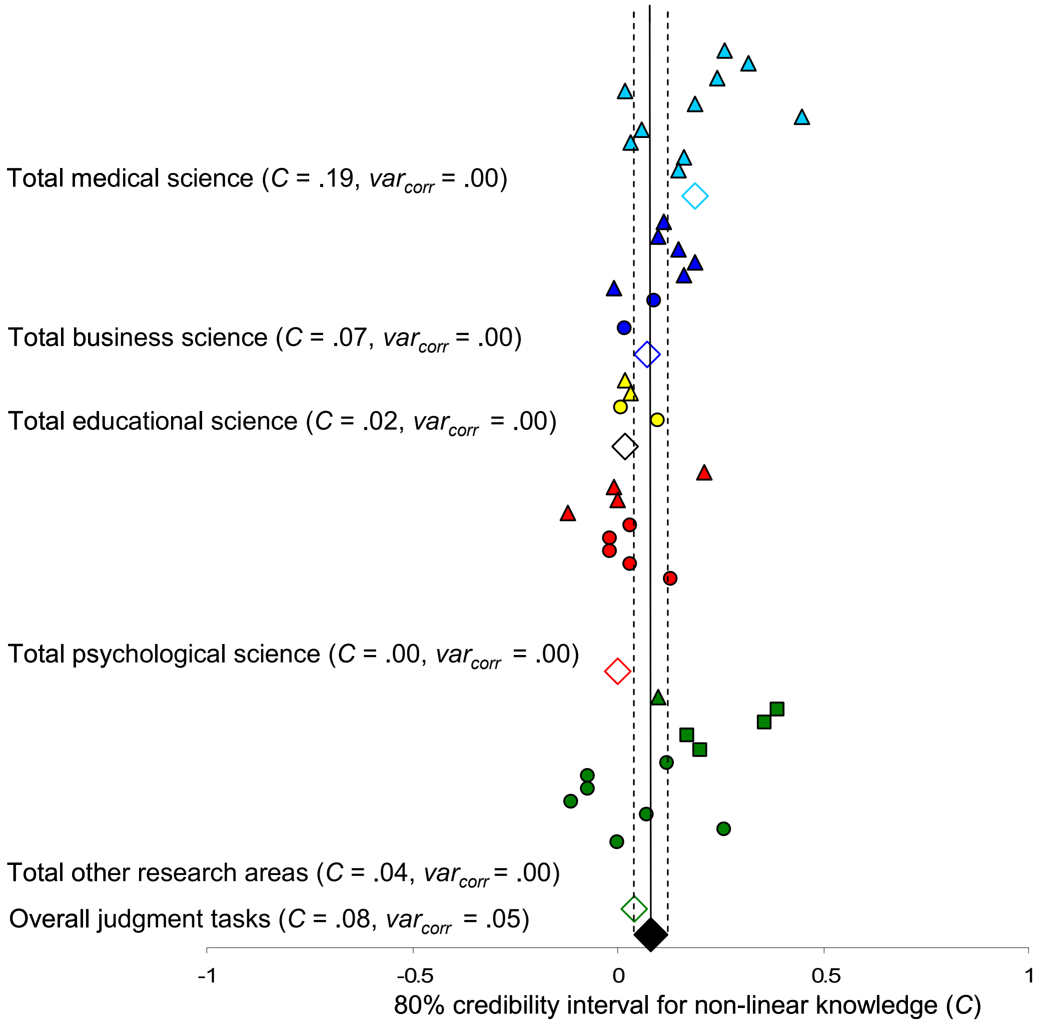
Table 23

Bare-bones meta-analysis according to the method of Hunter-Schmidt (2004), supplemented by a trim-and-fill analysis of the task-predictability component (R_e), separated into research area and experience level

Research area	*k*	*N*	R_e	var_{corr}	80% CI		75%
Medicine	10	258	.67	.00	.67	.67	105.89
Business	9	239	.71	.02	.53	.89	34.97
Education	4	156	.70	.00	.70	.70	257.26
Psychology	14/16	249/265	.68/.62	.00/.05	.59/.33	.77/.91	77.79/32.53
Miscellaneous	12/16	249/289	.88/.82	.01/.07	.76/.07	1.00/1.16	23.75/12.82
Overall	49/58	1151/1348	.73/.58	.01/.12	.59/.13	.88/1.00	44.21/14.37
Experts in:							
Business	6/8	116/133	.62/.50	.00/.07	.62/.14	.62/.85	108.29/34.31
Education	2	40	.68	.00	.68	.68	1690.13
Psychology	4/5	59/76	.80/.61	.00/.09	.80/.22	.80/1.00	256.36/22.98
Miscellaneous	5/7	15/23	.69/.34	.00/.00	.69/.33	.69/.33	356.44/153.73
Overall[a]	27/32	488/540	.68/.58	.00/.05	.68/.28	.68/.88	126.13/36.47
Students in:							
Business	3	123	.79	.02	.66	.97	13.91
Education	2	116	.71	.00	.71	.71	145.93
Psychology	9/13	176/220	.63/.52	.00/.08	.58/.14	.69/.89	91.12/27.88
Miscellaneous	11/14	234/267	.89/.79	.00/.06	.81/.47	.97/1.12	39.67/12.57
Overall	26/32	663/787	.77/.61	.02/.13	.60/.14	.94/1.00	31.23/12.10

Note. *k* = number of correlations (i.e. judgment tasks). *N* = total sample size for all judgment tasks combined. R_e = weighted mean correlation according to Hunter and Schmidt (2004). var_{corr} = corrected variation according to Hunter and Schmidt (2004, variance of true score correlation). 80% CI = 80% credibility interval for true score correlation distribution. 75% = Percentage variance of observed correlations due to all artefacts, if below 75%, it indicates moderator variable. /Results of the fill-and-trim analyses after a publication bias is indicated. [a]this analysis includes also medical experts. Grey boxes: Results not confirmed by the trim-and-fill analysis.

Unmodeled knowledge. In contrast to other components of the *LME*, the overall average value for the unmodeled knowledge component *C* was quite low (*C* = .08), corresponding to an r_c^2 value of only .16% (see Figure 17 and Appendix G: Table 1). Furthermore, there is no variation in the data. Hence, we also reran our analysis, separating our data into different research areas as well as by experience level within research areas. Finally, our *C* component analysis was completely confirmed by our publication-bias analysis supplemented with the trim-and-fill method. To summarize: All values remain low, with a small variance, and indicate no moderator variables.



You will find the legend on page 119.

Figure 17. The forest plot of the non-linear knowledge component (C), separated into the applied research areas, and within these by experience level. The studies in the forest plots are in the same order as in Table 5 and 6.

5.2.2 Psychometric meta-analysis

In the following, our results of a psychometric meta-analysis will be described. For an overview of our correction we refer to chapter 4.5.2.

We must mention that we only corrected this analysis for measurement error, and the study by Szucko and Kleinmuntz (1981) for the artefact of dichotomisation.

5.2.2.1 Judgment achievement

As noted previously, in educational, psychological, and the miscellaneous area, there were no area-specific retest-reliability values available for our measurement-error correction. However, in line with Hunter and Schmidt (2004), we assume that there is no perfect study and that there are always artefacts such as measurement error included. Hence, we used three different retest-reliability values for our measurement-error corrections: the average retest-reliability value of .78 by Ashton (2000), supplement by two extreme values (.90 as a high retest-reliability value, and .50 as a low retest-reliability value).

The psychometric meta-analysis is summarized in Table 24 and in Figure 18. There is a moderate mean (.45) from the 49 achievement correlations across 1151 judgments, if we assume high retest-reliability values (.90). This result is supported by our correction using .78 as retest-reliability value. On the other hand, if we assume a retest-reliability value of .50, judgment achievement clearly increases to a high value. In addition, the 75% rule no longer indicates any moderator variables and presents the value of measurement-error corrections. Although the correction allows only a small increase in the explanation of the judgment achievement variance when it’s corrected by .90 retest-reliability value, so it's clearly an increase by a .50 retest-reliability correction and also shows how different retest-reliability values influence our results. Without such an analysis, we would clearly underestimate the value of the average judgment-achievement value.

However, as the 75% rule also indicates moderator variables, we checked for differences in research areas. Hence, we reran our analysis, separating it into five suggested research areas. This meta-analysis lead to an increase of judgment achievement in medical science, from a moderator to a high value (.53). In this category, additionally 13% variance in observed correlation is attributable to all artefacts explained by measurement error. The highest increase in the explained variance is found in educational science (178%). In the other areas there is no increase in the explanation of the variance by our measurement-error correction. All judgment achievement increases to a high value, except in psychology science, if we assume retest-reliability values of .78 in our analysis. In psychology science, no high value is reached even if we assume a retest-reliability value of .50. Finally, our analysis also indicated moderator variables in business and in other research areas, so, we reran the analysis, separating the experience level within the areas.

First, our analysis across the areas, separating the experience level, shows that experts reach a slightly higher judgment achievement than students and indicates no moderator variables. Hence, the analysis separating experts within the research areas clearly shows that there are no moderator variables evident in any research area. On the other hand, this analysis still reveals moderator variables in the category of business-science students. A closer look at the scatter plots of students' judgment achievement in business reveals that one study (Wright, 1979) had low values and could influence our results. If we exclude this study, judgment achievement increases (r_a = .97; var_{corr} = .00; k = 2; N = 76), but still indicates moderator variables (75% rule = 30.51).

Finally, our trim-and-fill application if a publication bias is indicated confirms the analysis done in the bare-bones meta-analysis. The results can be found in parentheses in Table 24, below the psychometric meta-analysis results. However, there was one exception: The judgment achievement of students in business science deceases to a moderate value instead of a high value.

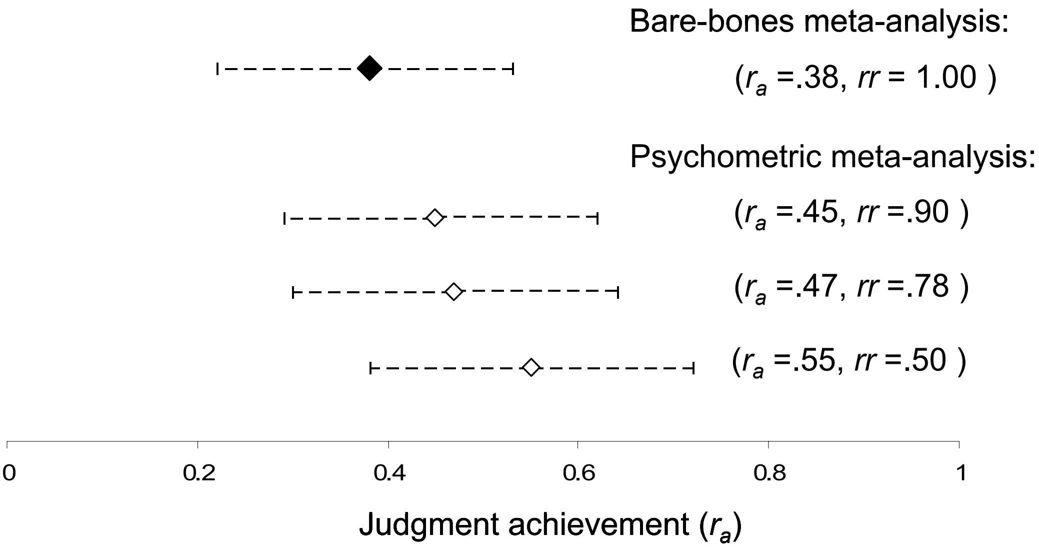


Figure 18. A comparison of the different corrected psychometric analyses.

Table 24

Psychometric meta-analysis of judgment achievements (r_a) in different research areas, separated by experience levels

				Overall			*Experts*			*Students*		
Research area	*rr*	*k (experts)*	*N (experts)*	r_a	var_{corr}	75%	r_a	var_{corr}	75%	r_a	var_{corr}	75%
Medical science		10	258	**.53**	**.00**	**170.93**	**.53**	**.00**	**170.93**	[a]	[a]	[a]
				(.53)	(.00)	(142.25)	(.53)	(.00)	(142.25)			
Business science		9(6)	239(116)	**.55**	**.09**	**24.45**	**.40**	**.00**	**87.73**	**.70**	**.11**	**8.52**
				(.22)	(.31)	(13.56)	(.27)	(.05)	(60.24)			
Education science	.90	4(2)	156 (40)	**.51**	**.00**	**355.11**	**.62**	**.00**	**97569**	**.55**	**.00**	**82558**
				(.41)	(.02)	(74.99)				(.36)[b]	(.00)[b]	(27136)[b]
	.78			**.58**	**.00**	**347.80**	**.72**	**.00**	**97569**	**.59**	**.00**	**82558**
				(.46)	(.02)	(73.97)				(.40)[b]	(.00)[b]	(27137)[b]
	.50			**.82**	**.00**	**349.47**	--	--	--	**.73**	**.00**	**82558**
				(.67)	(.05)	(74.25)				(.55)[b]	(.00)[b]	(31634)[b]
Psychology	.90	14(4)	249(59)	**.24**	**.00**	**448.54**	**.11**	**.00**	**975.77**	**.29**	**.00**	**607.07**
				(.23)	(.00)	(319.78)	(.06)	(.00)	(635.55)			
	.78			**.27**	**.00**	**449.50**	**.12**	**.00**	**975.77**	**.32**	**.00**	**608.58**
				(.26)	(.00)	(320.21)	(.07)	(.00)	(635.55)			
	.50			**.39**	**.00**	**457.90**	**.14**	**.00**	**975.77**	**.46**	**.00**	**626.40**
				(.36)	(.00)	(325.32)	(.09)	(.00)	(635.55)			
Miscellaneous	.90	12(5)	249(15)	**.49**	**.02**	**67.55**	**.68**	**.00**	**401.61**	**.48**	**.00**	**86.55**
				(.36)	(.08)	(43.98)	(.31)	(.00)	(158.46)	(.35)	(.07)	(53.59)
	.78			**.52**	**.02**	**67.78**	[c]	[c]	[c]	**.55**	**.01**	**86.55**
				(.42)	(.11)	(44.02)				(.42)	(.11)	(44.02)
	.50			**.86**	**.06**	**71.08**	[c]	[c]	[c]	**.85**	**.02**	**86.50**
				(.64)	(.26)	(44.58)				(.64)	(.26)	(44.48)
Overall	.90	49 (27)	1151 (488)	**.45**	**.02**	**74.55**	**.47**	**.00**	**135.15**	**.46**	**.02**	**64.20**
				(.40)	(.11)	(44.80)	(.41)	(.01)	(86.93)	(.40)	(.07)	(42.13)
	.78			**.48**	**.02**	**74.79**	**.49**	**.00**	**134.66**	**.51**	**.03**	**64.26**
				(.37)	(.10)	(37.73)	(.42)	(.01)	(86.71)	(.45)	(.09)	(42.13)
	.50			**.55**	**.02**	**82.46**	**.53**	**.00**	**139.80**	**.72**	**.05**	**71.00**
				(.52)	(.19)	(46.32)	(.47)	(.01)	(89.30)	(.64)	(.17)	(45.23)

Note. Values enclosed in parentheses represent our results of the trim-and-fill method application if a publication bias is indicated. *rr* = retest-reliability values used in our measurement-error corrections. *k* = Number of correlations according to Hunter and Schmidt (2004). *N* = Total sample size according to Hunter and Schmidt (2004). r_a = mean true score correlation according to Hunter and Schmidt (2004). var_{corr} = corrected variation according to Hunter and Schmidt (2004, variance of true score correlation). 75% = Percentage variance of observed correlations due to all artefacts, if below 75%, it indicates moderator variable. -- mean true score correlation increased the value of 1. [a]In medical science only experts are included. [b]we reran the analysis and substituted the .09 value with a .90 value. [c]no further correction because only meteorologists in this category are included (*rr* = .93, see Ashton, 2000). Grey boxes: Results not confirmed by our trim-and-fill analysis.

5.2.2.2 Judgment achievement components

Knowledge component. In our psychometric meta-analysis, the knowledge component across areas increased minimally at 10% (*G* = .77) to 15% (*G* = .83, see Table 25). Hence, we revealed an increase of 1.57% to 9.37% attributable for measurement errors due to all artifacts. In addition, the 75% rule clearly indicates that there were true differences in effect size across judgment tasks. Accordingly, separate meta-analyses were calculated for research areas. This analysis shows that our increase in the *G* component in this meta-analysis across research areas is dominated by the medical category. In medical science, the knowledge component increases clearly to .82, and also the percent variance in observed correlation attributable to all artifacts increased to 3.66%. However, there is still a moderate knowledge component (.42) in psychological science. After a measurement correction with a retest-reliability value of .50, also the knowledge component in psychological science increases to a high level (.76). Finally, there are still moderator variables indicated in all areas except educational science. Hence, we reran our analysis, separating experience levels. Against our expectation, our analysis implied that experts have lower knowledge components than students across research areas, but both values are still high. However, our analysis also reveals that there are differences in research areas. Hence, the experience level could explain the heterogeneity in psychology and other research areas. In education and other research areas, the experts also have higher knowledge components than students, but not in psychology science. In psychology science, there is a clear difference between low experts' knowledge components to high students' knowledge components. In addition, in business science, the same pattern – that students have a better knowledge component – is revealed, but both *G* components in business science are still high, and moderator variables are indicated in both categories.

Finally, our trim-and-fill application if a publication bias is indicated confirmed our previous analysis for a bare-bones meta-analysis.

Table 25

Psychometric meta-analysis of the linear knowledge component (G) in different research areas, separated by experience levels

				Overall			*Experts*			*Students*		
Research area	*rr*	*k (experts)*	*N (experts)*	*G*	var_{corr}	75%	*G*	var_{corr}	75%	*G*	var_{corr}	75%
Medical science		10	258	**.82** (.81)	**.02** (.04)	**68.47** (53.59)	**.82** (.81)	**.02** (.04)	**68.47** (53.59)	[a]	[a]	[a]
Business science		9(6)	239(116)	**.73**	**.08**	**15.85**	**.60** (.50)	**.06** (.13)	**35.56** (26.45)	**.86**	**.06**	**6.95**
Education science	.90	4(2)	156(40)	**.97**	**.00**	**452.80**	**.98**	**.00**	**313.80**	--	--	--
	.78			--	--	--	--	--	--	--	--	--
	.50			--	--	--	--	--	--	--	--	--
Psychology	.90	9(4)	105(59)	**.42** (.27)	**.03** (.13)	**73.77** (45.46)	**.18** (.13)	**.00** (.00)	**444.93** (302.49)	**.72** (.41)	**.04** (.21)	**57.65** (35.14)
	.78			**.48** (.31)	**.04** (.17)	**73.77** (45.46)	**.19** (.15)	**.00** (.00)	**444.93** (302.49)	.82 (.47)	**.05** (.28)	**57.65** (35.14)
	.50			**.76** (.48)	**.09** (.42)	**73.77** (45.46)	**.24** (.18)	**.00** (.00)	**444.93** (302.49)	--	--	--
Miscellaneous	.90	12(5)	249(15)	**.75** (.54)	**.08** (.20)	**19.27** (17. 90)	**.96** (.74)	**.00** (.03)	**768.55** (80.55)	**.74** (.60)	**.08** (.16)	**24.15** (22.03)
	.78			**.87** (.62)	**.10** (.26)	**19.61** (17. 95)	[b]	[b]	[b]	**.85** (.69)	**.10** (.21)	**24.15** (22.03)
	.50			--	--	--	[b]	[b]	[b]	--	--	--
Overall	.90	44(27)	1007(488)	**.77** (.75)	**.07** (.08)	**37.11** (34.24)	**.71** (.65)	**.06** (.10)	**49.80** (41.32)	**.77** (.57)	**.06** (.20)	**22.00** (14.44)
	.78			**.83** (.79)	**.08** (.08)	**35.28** (33.42)	**.75** (.68)	**.06** (.11)	**48.62** (40.61)	**.88** (.65)	**.08** (.26)	**21.86** (14.42)
	.50			--	--	--	**.85** (.80)	**.07** (.14)	**56.89** (45.74)	--	--	--

Note. Values enclosed in parentheses represent our results of the trim-and-fill method application if a publication bias is indicated. *rr* = retest-reliability values used in our measurement error corrections. *k* = Number of correlations according to Hunter and Schmidt (2004). *N* = Total sample size according to Hunter and Schmidt (2004). *G* = mean true score correlation according to Hunter and Schmidt (2004). var_{corr} = corrected variation according to Hunter and Schmidt (2004, variance of true score correlation). 75% = Percentage variance of observed correlations due to all artefacts, if below 75%, it indicates moderator variable.-- mean true score correlation increased the value of 1. [a]In the medical science only experts are included. [b]no further correction because only meteorologists are included in this category (*rr* = .93, see Ashton, 2000). Grey boxes: Results not confirmed by the trim-and-fill analysis.

Consistency components. Our measurement-corrected consistency component increases minimally at 8%, if we assume a high retest-reliability value of .90 in our correction (see Table 26). In comparison to our bare-bones meta-analysis, this correction procedure reveals that there are now no moderators indicated.

However, we reran our analysis, separating different research areas. Hence, in comparison to our bare-bones analysis of the consistency component, our medical-science and educational-science consistency analysis reveals no further moderator variables. However, in business and also in psychology science there are still moderator variables indicated in the constancy component. Hence, we analysed the data considering also the experience level of the judges within the research areas. At first glance, our results are in line with our expectation that expert's judge more consistently than students. Although there are no moderator variables indicated in the different experience levels, we reran our analysis considering also the research areas, in order to check if there are clearly no differences in research areas. Our analysis shows that although the experts' and also the students' analysis indicated no moderator variables, there are clearly variables indicated in psychological experts' and in students' business consistency.

Finally, our trim-and-fill analysis used if a publication bias is indicated reveals that maybe moderator variables are also indicated in our across analysis, as implied by our bare-bones meta-analysis. In addition, also the educational areas now indicated moderator variables. Furthermore, all found publication bias could not be eliminated by the experience level of the judges. It must, however, be mentioned that all consistency components after a trim-and-fill analysis still remain high.

Table 26

Psychometric meta-analysis of the consistency component (R_s) in different research areas, separated by experience levels

				Overall			*Experts*			*Students*		
Research area	*rr*	*k (experts)*	*N (experts)*	R_s	var_{corr}	75%	R_s	var_{corr}	75%	R_s	var_{corr}	75%
Medical science		10	258	**.96** (.92)	**.00** (.00)	**126.87** (84.85)	**.96** (.92)	**.00** (.00)	**126.87** (84.85)	[a]	[a]	[a]
Business science		9(6)	239(116)	**.89**	**.02**	**28.60**	**.69** (.67)	**.00** (.00)	**108.29** (105.37)	**.85**	**.03**	**12.68**
Education science	.90	4(2)	156(40)	**.93** (.63)	**.00** (.18)	**554.87** (23.86)	**.96**	**.00**	**1241**	--	--	--
	.78			**.96** (.67)	**.00** (.20)	**533.78** (22.80)	--	--	--	--	--	--
	.50			--	--	--	--	--	--	--	--	--
Psychology	.90	12(4)	150(59)	**.83**	**.01**	**71.34**	**.89** (.84)	**.01** (.02)	**48.83** (33.04)	**.78** (.59)	**.00** (.09)	**107.28** (35.15)
	.78			**.89**	**.01**	**71.34**	**.96** (.91)	**.01** (.03)	**48.83** (33.04)	**.84** (.60)	**.00** (.09)	**107.28** (35.15)
	.50			--	--	--	--	--	--	--	--	--
Miscellaneous	.90	12(5)	249(15)	**.75** (.67)	**.00** (.05)	**90.92** (34.01)	**.98** (.78)	**.00** (.06)	**1724** (66.77)	**.73** (.66)	**.00** (.08)	**148.50** (79.68)
	.78			**.80** (.72)	**.00** (.06)	**93.00** (34.35)	[b]	[b]	[b]	**.79** (.72)	**.00** (.09)	**148.50** (79.68)
	.50			**.97** (.88)	**.00** (.09)	**123.10** (37.66)	[b]	[b]	[b]	**.98** (.89)	**.00** (.01)	**148.50** (79.68)
Overall	.90	47(27)	1052(488)	**.85** (.66)	**.00** (.14)	**100.63** (17.31)	**.92** (.90)	**.00** (.01)	**138.60** (52.26)	**.78** (.61)	**.00** (.11)	**139.48** (24.43)
	.78			**.88** (.69)	**.01** (.15)	**94.73** (16.93)	**.93** (.92)	**.00** (.01)	**132.60** (50.06)	**.83** (.65)	**.00** (.13)	**129.98** (24.10)
	.50			**.99** (.79)	**.00** (.19)	**137.75** (19.85)	**.98** (.97)	**.00** (.00)	**209.62** (79.27)	**.99** (.79)	**.00** (.19)	**146.02** (24.93)

Note. Values enclosed in parentheses represent our results of the trim-and-fill method application if a publication bias is indicated. *rr* = retest-reliability values used in our measurement error corrections. *k* = Number of correlations according to Hunter and Schmidt (2004). *N* = Total sample size according to Hunter and Schmidt (2004). R_s = mean true score correlation according to Hunter and Schmidt (2004). var_{corr} = corrected variation according to Hunter and Schmidt (2004, variance of true score correlation). 75% = Percentage variance of observed correlations due to all artefacts, if below 75%, it indicates moderator variable. -- mean true score correlation increased the value of 1. [a]In medical science only experts are included. [b]no further correction because only meteorologists are included in this category (*rr* = .93, see Ashton, 2000). Grey boxes: Results not confirmed by the trim-and-fill analysis.

The environmental predictability component. In our measurement corrections including the environmental component, we only used a one-side correction, hence, this correction is a conservative one. If we look at the environmental-component analysis across areas, however, there is also a high value, which increases by 9% (see Table 27). In addition, there are clearly moderator variables indicated. However, a rerun of our analysis, separating different research areas, reveals that this moderator variable indication is dominated by study of the other research category. On the other hand, if we rerun the analysis and separate the experience level, it is also clear, that students' tasks indicated moderator variables. Hence, we reran an analysis, separating the experience level within the research areas. However, this analysis is also in line with our bare-bones meta-analysis, revealing the same pattern that a meta-analysis of task in psychology or other research areas done by students still indicated moderator variables.

To summarise: In our psychometric meta-analysis of the environmental-predictability component, the same pattern as in our bare-bones meta-analysis was found. However, our corrections let clearly increase the explained heterogeneity and also the task predictability values. Finally, our publication-bias analysis with the trim-and-fill method confirms the analysis done in the bare-bones meta-analysis. It must be mentioned that in the environmental-predictability analysis reached a value up to 1 are excluded to prevent an overcorrection.

The C Component. As the component *C* is almost zero and indicates no heterogeneity, we added our psychometric meta-analysis in the Appendix G: Table 2.

Table 27

Psychometric meta-analysis of the environmental-predictability component (R_e) in different research areas, separated by experience level

				Overall			*Experts*			*Students*		
Research area	*rr*	*k (experts)*	*N (experts)*	R_e	var_{corr}	75%	R_e	var_{corr}	75%	R_e	var_{corr}	75%
Medical science		10	258	**.92**	**.00**	**198.55**	**.92**	**.00**	**198.55**	[a]	[a]	[a]
Business science		9(6)	239(116)	[b]	[b]	[b]	[b]	[b]	[b]	[b]	[b]	[b]
Education science	.90	4(2)	156(40)	**.74**	**.00**	**257.26**	**.72**	**.00**	**1690**	**.75**	**.00**	**145.93**
	.78			**.80**	**.00**	**257.26**	**.77**	**.00**	**1690**	**.80**	**.00**	**145.93**
	.50			**.99**	**.00**	**257.26**	**.96**	**.00**	**1690**	--	--	--
Psychology	.90	14(4)	249(59)	**.72** (.64)	**.00** (.05)	**78.62** (32.73)	[b]	[b]	[b]	**.68** (.54)	**.00** (.09)	**85.35** (27.90)
	.78			**.74** (.67)	**.00** (.06)	**83.82** (34.04)	[b]	[b]	[b]	**.73** (.58)	**.00** (.11)	**85.35** (28.25)
	.50			**.82** (.75)	**.00** (.06)	**134.00** (46.33)	[b]	[b]	[b]	**.91** (.67)	**.00** (.14)	**85.35** (32.57)
Miscellaneous	.90	12(5)	249(15)	**.93** (.82)	**.00** (.07)	**23.74** (12.82)	[b]	[b]	[b]	**.94** (.83)	**.00** (.07)	**39.67** (12.57)
	.78			--	--	--	[b]	[b]	[b]	--	--	--
	.50			--	--	--	[b]	[b]	[b]	--	--	--
Overall	.90	49(27)	1151(488)	**.81** (.64)	**.00** (.14)	**66.00** (16.59)	**.76** (.59)	**.00** (.06)	**157.52** (36.54)	**.81** (.64)	**.02** (.14)	**31.45** (12.11)
	.78			**.83** (.66)	**.01** (.15)	**61.04** (16.30)	**.78** (.59)	**.00** (.06)	**155.15** (36.58)	**.85** (.79)	**.02** (.21)	**35.38** (17.09)
	.50			**.94** (.72)	**.01** (.18)	**88.00** (18.84)	**.80** (.59)	**.00** (.06)	**170.79** (36.87)	--	--	--

Note. Values enclosed in parentheses represent our results of the trim-and-fill method application if a publication bias is indicated. *rr* = retest-reliability values used in our measurement-error corrections. *k* = Number of correlations according to Hunter and Schmidt (2004). *N* = Total sample size according to Hunter and Schmidt (2004). R_e = mean true score correlation according to Hunter and Schmidt (2004). var_{corr} = corrected variation according to Hunter and Schmidt (2004, variance of true score correlation). 75% = Percentage variance of observed correlations due to all artefacts, if below 75%, it indicates moderator variable. -- mean true score correlation increased the value of 1. [a]In medical science only experts are included. [b]see bare-bones meta-analysis, no correction because this category includes only objective criterions. Grey boxes: Results not confirmed by the trim-and-fill analysis.

5.2.3 Intercorrelation of the components

The intercorrelations of the *LME* components across areas reveal a strong correlation between judgment achievement and all components, except for the moderate correlation between judgment achievement and the component *C* (see Table 28). Even though this is the strongest *C* component correlation, the *C* component correlates only weakly with the other components. The same pattern is found with the R_s component.

However, if we separate the data base into different experience levels, such a strong correlation between judgment achievement and judgment consistency in experts' judgments is not confirmed. In addition, there is a clear increase in the *C* component correlations with other components in experts' judgments except for the R_s component. Furthermore, also the R_e component and R_s component now reach a strong correlation. On the other hand, *LME* intercorrelation of the components in students' judgments reveals a low correlation between the r_a and *C* component.

In addition, the intercorrelation between the components in different research areas of experts' judgment confirms a heterogeneous picture of the *LME* intercorrelations (see Appendix G: Tables 3, 4). First of all, in the other research area there are overall components a high intercorrelation between the components – it must be mentioned – that this category only includes Stewart's meteorology studies (1990, 1997). Secondly, we also found a highly negative correlation in psychology science, whereas judgment achievement is strongly negatively correlated with the R_e and R_s components. Thirdly, high negative correlations are also found in education science between the *G* components and all others components. On the other hand, if we look at students' judgments, the mentioned patterns are not confirmed.

Altogether, beside the found heterogeneity in our *LME* components, there is also a great heterogeneity in the intercorrelation between the *LME* components (see Tables 28 and 29; Appendix G:

Tables 3, 4). This should be taken into account in the interpretation of our results.

Table 28

Intercorrelation of the LME components

	Components				
Components	r_a	G	R_s	R_e	C
r_a	--	.81**	.57**	.56**	.44**
G	.81**	--	.22	.50**	.18
R_s	.57**	.22	--	.29*	.27
R_e	.56**	.50**	.29*	--	.17
C	.44**	.18	.27	.17	--
Experts:					
r_a	--	.92**	.41*	.65**	.63**
G	.92**	--	.34	.45*	.48*
R_s	.41*	.34	--	.50**	.03
R_e	.65**	.45*	.50**	--	.35
C	.63**	.48*	.03	.35	--
Students:					
r_a	--	.66**	.84**	.51**	.29
G	.66**	--	.40	.49*	-.20
R_s	.84**	.40	--	.50**	.35
R_e	.51**	.49*	.50**	--	.19
C	.29	-.20	.35	.19	--

Note. ** Correlation is significant at the .001 level (2-tailes).

* Correlation is significant at the .005 level (2-tailes).

Table 29

Intercorrelation of the LME components in the different areas

Components in:	Components				
Medical science	r_a	G	R_s	R_e	C
r_a	--	.93**	.44	.90**	.63
G	.93**	--	.35	.73*	.40
R_s	.44	.35	--	.53	.29
R_e	.90**	.73*	.53	--	.67*
C	.63	.40	.29	.67*	--
Business science					
r_a	--	.92**	.54	.97**	-.33
G	.92**	--	.43	.91**	-.02
R_s	.54	.43	--	.47	-.66
R_e	.97**	.91**	.47	--	-.21
C	-.33	-.02	-.66	-.21	--
Education science					
r_a	--	.86	.79	-.63	-.49
G	.86	--	.37	-.93	-.86
R_s	.79	.37	--	-.05	15
R_e	-.63	-.93	-.05	--	.96*
C	-.49	-.86	.15	.96*	--
Psychology science					
r_a	--	.45	-.19	-.07	.25
G	.45	--	-.76*	.45	-.07
R_s	-.19	-.76*	--	.29	-.27
R_e	-.07	.45	.29	--	-.65
C	.25	-.07	-.27	-.65	--
Miscellaneous					
r_a	--	.78**	.85**	.40	.46
G	.78**	--	.74**	-.10	-.05
R_s	.85**	.74**	--	-.08	.48
R_e	.40	-.10	-.08	--	.44
C	.46	-.05	.48	.44	--

Note. ** Correlation is significant at the .001 level (2-tailes).

* Correlation is significant at the .005 level (2-tailes).

5.2.4 Robustness analysis

To control for the robustness of our reported results, we checked several factors such as a) the type of models such as the fixed-effects models or random-effect model (see chapter 4.4.4) and b) the used weighting strategy (see chapter 4.5.1.1) and finally, c) the type of correlation (Product-non product correlation, see chapter 4.3.7).

5.2.4.1 Type of used model

To check if our results depend on the used random-effect model (see Figure 19, random-effect model, HS = Hunter and Schmidt estimator) we checked our bare-bones meta-analysis results (see Tables 20 - 23) against a fixed-effect model and a further random-effect model with the estimator suggested by DerSimonian and Laird (*DL*, 1986). An overview of the overall judgment achievement estimations dependent on different models is found in Figure 19 (for details see Appendix H: Tables 1 - 5). In addition, it is to mention that only in the Hunter and Schmidt estimation credibility intervals are used (represented by the dashed lines, see chapter 4.5.1.1). To summarize our analysis differs only slightly if we focus on all *LME* components overall and between research areas. These differences could be explained by rounding showing clearly that our results are robust. Hence, we conclude that our results are independent on the used models and assume that this is also the case within research areas.

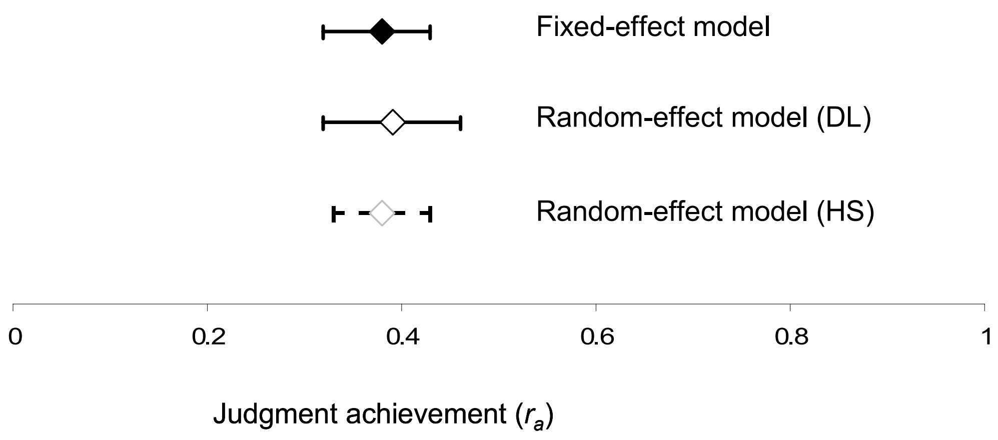


Figure 19. Comparison of different model used in meta-analytic research.

5.2.4.2 Weighting strategy

In the Table 30 our meta-analysis are repeated for all *LME* components across subject areas and between subject areas. However, we used another weighting strategy with the number of judges and profiles, representing the used weighting strategy we used in our meta-analysis with idiographic data. If we compare this analysis with our bare-bones meta-analysis considering also nomothetic data then it's clearly that the values are comparable with some exceptions as the moderate achievement values in medical (.40) decreases to a low value (.29). However, as the variance increases also the level of moderator variables increases leading to more moderator variables indications. Hence, our analysis of the nomothetic data base is more conservative than the estimation of our idiographic data base.

Table 30

Weighting strategy judges and profiles: Descriptive statistics for the components of correlations of the LME according to the method of Hunter-Schmidt (2004)

		r_a			G			R_s			R_e			C		
Research area:	k	M	var_{corr}	75%	M	var_{corr}	75%	M	var_{corr}	75%	M	var_{corr}	75%	M	var_{corr}	75%
Medical science	10	.29	.03	2.3*	.47	.07	.65*	.80	.01	.96*	.58	.01	2.7*	.12	.01	6.3*
Business	9	.46	.10	38*	.56	.11	.29*	.78	.01	.66*	.64	.03	.69*	.09	.00	24.4*
Education	4	.39	.01	1.45*	.73	.02	.28*	.73	.01	.35*	.71	.00	2.06*	.02	.00	27.03*
Psychology	11	.22	.01	7.0*	.48	.11	.95*	.87	.00	1.88*	.74	.02	1.40*	.00	.01	22.96*
Miscellaneous	12	.55	.06	5*	.75	.06	.00*	.77	.02	.75*	.89	.01	2.01*	.09	.02	4.28*
Overall	46[a]	.40	.05	.78*	.64	.08	29*	.77	.01	.72*	.73	.03	.53*	.07	.01	6.8*

Note. k = Number of correlations according to Hunter and Schmidt (2004). var_{corr} = corrected variation according to Hunter and Schmidt (2004, variance of true score correlation). 75% = Percentage variance of observed correlations due to all artefacts, if below 75%, it indicates moderator variable. *Moderator variables indicated. [a]The study by Reynolds and Gifford (2001) is excluded representing three tasks.

5.2.4.3 Type of correlation

For this robustness analysis we refer to Kaufmann and Athanasou (2009). To summarize it, with a reduced sample base we found no indication that this sample the type of correlations systematically influenced the robustness of our results.

5.2.4.4 Conclusion

Our different sensitivity analysis shows clearly that our results are robust against the introduced factors such as used model and the type of correlation. On the other hand, we would like to highlight that the used weighting strategy using also profiles leads to liberal results and should therefore, also be considered in the interpretation of our results.

6 DISCUSSION

Brunswik's achievement concept formalized by the *LME* led to numerous publications revealing why some people are more accurate than others. Hence, these meta-analysis gives an impression of how well artifact corrected judges are in different areas separated by experience level. To overcome the ecological and the individualistic fallacy, two types of data base were used. The idiographic data base includes only *LME* component by single judges. This data base is supplemented by *LME* component across judges in our nomothetic data base.

In the following, we will first discuss specific aspects in relation to our data base, starting with the idiographic data base, and then follow-up with specific aspects concerning also our nomothetic data base. Secondly, we will also focus on the limitations of our meta-analyses. Finally, we will give a comprehensive outlook for further analyses and studies.

6.1 Idiographic-based meta-analysis

To overcome the ecological fallacy (Robinson, 1950, see chapter 2.3.4.1), we conducted a meta-analysis based on individual-level data.

The major finding of the meta-analysis applied to all selected studies is that humans judge a given criterion with a moderate achievement of .38 (Figure 9, Table 13). Additionally, one has to take into account the large credibility interval of the judgment achievement. Furthermore, there are individual differences among the judges' achievements. However, all credibility intervals are higher than zero, hence, the relationship generalization across persons in our meta-analysis is supported. But, can a similar conclusion be drawn considering different domains? Better judgment achievements were attained by those studies applied to the research areas denoted here as "other" and to the educational research areas (Table 13). Studies applied to the medical, business, or psychology sciences showed lower judgment achievements.

To clarify the contrasting results of the above judgment achievements, their components were considered. Firstly, judgment achievement can increase from a moderate level of .38 (judgment achievement) to a high level (.55, error-free judgment achievement, see Table 16) under optimal situations – except for those studies applied to business, even if they are also corrected for artefacts.

Furthermore, the environmental predictability is related to the moderate achievement level. In addition, moderate judgment achievement is related to the high level of consistency that was attained in both the combined and separated studies' meta-analysis.

We would like to highlight that knowledge introduces the highest variability, which should be checked also with our nomothetic data base. In addition, further analysis reveals possible moderator variables, such as experience levels within research areas. It was surprising that Shanteaus (2002) recommendation that experts reach a better judgment achievement was not confirmed in business and psychology science. However, as the data base is too small, we refer to our discussion on our nomothetic data base. In addition, the obtained results suggest that judgment achievement is also influenced by other factors, such as the number of cues, as the exclusion of the study with the highest number of cues lead to a higher knowledge value. As Miller (1956) showed, memory limits the amount of information or number of cues that can be processed. It can be argued that the subjects judged tasks with a limited number of cues more accurately than tasks with more cues. The effect of increasing information by adding cues has been addressed by several researchers (Nystedt, 1974; Nystedt & Magnusson, 1972).

To summarize, our results lead us to the conclusion that there are some area differences. Especially given that the differences between areas still remain after artefact correction. Therefore, we assume that the different research areas could be a possible moderator variable also in lens studies. In addition, because of the underlying heterogeneity of the *LME* components, our results imply that research theories that mainly focus on the task-side (see the Fast and Frugal Heuristic Approach, Gigerenzer et al., 1999) or focus on the judge-side, clearly short cut the explanations of differences in judgment achievement. In line with Brunswik, we therefore recommend to focus on both sides – after the data is corrected for possible artefacts – in order to find the main sources of the judgment achievement heterogeneity on the task and judges' side.

In addition, it is surprising that although an array of reviews on judgment achievement have been published (e.g. Aegisdottir et al., 2006; Grove et al., 2000, see chapter 2), none of them focus on idiographic values. So, against our intention, our results could not be compared directly with other results found, and, we have to refer to the discussion of our nomothetic data base.

However, as with most research, there are limitations to consider in the interpretation of the results found in the reported analysis. Hence, the six major limitations are illustrated in the following. The six points are related to the *Hunter-Schmidt approach*, *artefact corrections*, *cue-intercorrelations*, *publication bias*, and further *robustness analysis*. Additional information is also taken from Kaufmann et al., 2007.

First, it must be mentioned that the *Hunter-Schmidt method* is usually applied at the nomothetic level, i.e. across studies. Because the idiographic approach was used here, the analyses were carried out across persons. Therefore, a problem in our meta-analysis is that the judgment achievement of persons in the same study is more homogenous than that of persons between studies. In addition, the same persons judged two or more tasks; therefore the correlations were not independent in most cases. This problem was neglected, however, because of the size of the sample of persons (331 of 370) who judged only one task. Finally, the sample size of 30 judgment tasks and 370 analyzed judgment achievements that used an idiographic approach restricts the generality of the results. Therefore, the inclusion of studies using a nomothetic research approach led the generality of our results to increase. In addition, in our comparison between idiographic and nomothetic studies we clearly show that the idiographic approach is neglected in the current research. Hence, we would like to advise future researchers to reconsider the individual level, as this also prevents the ecological fallacy, as mentioned above. However, as far as we know, it is the first time that the Hunter-Schmidt method is used with individual data. There is no bare-bones nor psychometric meta-analysis published, according to Hunter and Schmidt (2004) using individual data. However, there is some theoretical discussion on this point in meta-analytic research (see Viechtbauer, 2007). In addition, in medical science, meta-analysis using individual patient data is highly recommended as an individualistic approach, but medical meta-analysis seldom uses a Hunter and Schmidt approach or prefers the fixed-effect model approach (e.g. Smith & Williamson, 2007). Altogether, our research brings back the focus to the individualistic approach recommended by Brunswik and ignored by most meta-analytic researchers.

We see this as a fruitful supplementation of the commonly used study data – to reveal the introduction of ecological fallacy and to focus on the aggregation problems (Wittmann, 1985, 1988). We highly recommend further research on this topic that could inspire research on meta-analytic approaches, and also, the reported evaluation research on meta-analytic research introduced in chapter 4.4.5. Hence, the reported point suggests a more idiographic-based psychological research instead of the dominance of the nomothetic research, as is typical in *JDM*. This claim for more idiographic studies was already supported in the early days by Brunswik.

Secondly, although in *our artefact correction* we used an idiographic approach, the corrected measurement values are nomothetic – individual data was, seldom available, only the authors of one study reported it. However, we do not assume that there is no measurement error included; hence, we used the nomothetic-based retest-reliability values. To prevent an overcorrection of our data, we did not correct any objective criterion for measurement error. Furthermore, we did not correct any cues' reliability values because of missing data. Consequently, our results are a rather conservative estimation.

Thirdly, to understand the eventual effect of the increase of the number of cues, we also have to take their *intercorrelations* with already existing cues into consideration, which was not possible in our analysis because of missing data.

Fourthly, there was no *publication bias* or trim-and-fill application used with this data base. As research on publication bias focuses only on aggregated data as study information, we see the current estimators as not suitable for our individual data base. In addition, as meta-analyses are seldom performed on individual data, we also promote further meta-analysis based on individual data, leading to more sophicated and suitable estimators for publication bias.

Fifthly, all *further robustness analyses* include our nomothetic-data-base. However, we see the separate analysis of the idiographic data also as a robustness analysis from the nomothetic-data-base point of view. Hence, one type of robustness analysis is the comparison of our idiographic and our nomothetic data base (for more information, see Kaufmann & Athanasou, 2009). Anyway, a comprehensive robustness analysis should urgently be done also with our idiographic data base.

In summary, the findings lead to the conclusion that humans predict a criterion with a moderate judgment achievement. Furthermore, the high error-free judgment achievement of persons implies that judgment achievement can be better than moderate. However, in line with Brunswik, we recommend the comparison of individual data first. As our results imply that judgment achievement is different between research areas, we further recommend comparing judgment achievement within one research area – separated by experience level – then the comparison with the values of other research areas (see Shanteau, 2002). Hence, our study clearly shows that the *LME*-based research is widely used in different research areas, and that this should be taken into account when studying judgment achievement.

As this data base includes 30 judgment tasks of 370 analysed individual sets of data, we added studies using a nomothetic data base, in order to overcome some of the reported limitations of our studies. By means of the increase of our data base, the validity of our results is checked.

6.2 Nomothetic-based meta-analysis

Our meta-analysis of the components of correlations of the *LME* of 31 studies incorporated 1055 persons in 49 judgment tasks. The major finding of our bare-bones meta-analysis is that humans' ability to judge a given criterion is moderate (.40). Our results are also supported by the previously discussed results of our idiographic data base and are in line with Karelaia and Hogarth's meta-analysis (2008).

However, to clarify the results of humans' ability for *JDM*, we looked at the underlying sources – the components – of judgment achievement. First, the moderate ability of humans' judgment achievement is also related to a high consistency of persons across judgment tasks. Second, the high environmental predictability shows that the criterion could be well judged. Therefore, the moderate ability of humans to judge is also related to high environmental predictability. Third, a high value of the error-free judgment achievement – the knowledge – of the judge is presented. Hence, the judgment achievement could increase from a moderate level of .40 to a high level of .69, except in studies applied to psychological science. The obtained results also support a good ability of humans' *JDM*, except for psychological judgments. However, it should be added that the results from our idiographic data based are confirmed that there is a great variability in our data, especially in the knowledge component.

In addition, our psychometric meta-analysis reveals that judgment achievement clearly increases because of artefact correction to a high level of .55, if we assume a .50 retest-reliability value for our correction. Hence, with the exception of psychological science – each research area finally reaches a high level. Our analysis also reveals a moderate *G* component in psychology science, in comparison to high values in all other areas. In addition, with an artefact-correction, each component clearly increases across research areas as well as between research areas.

Despite the exclusion of some studies, most meta-analyses still indicated moderating variables at the across-study level. However, the excluded studies are heterogeneous; therefore, no study characteristics which influence the components of achievement are revealed. If we compare our results with our idiographic data base, we find that the levels in business and psychological science are not the same. In business science, judgment achievement reaches a high value also in the nomothetic data base, without any measurement-error corrections, in comparison to the low values in our idiographic data base. The reverse pattern is found in psychological science. We would like to mention, however, the already introduced limitation of the number of judgment tasks in the idiographic data base leading to a reduced generalization of the results.

Consequently, because of our results based on our nomothetic data base, our suggested moderating variables, such as *experience level within research* areas, are discussed in more detail in the following. Against our expectation we found in both data bases that it is clearly not the case that experts judge better than non-experts, if we look at the overall level without any separation into research areas.

However, if we include the difference in areas in our experience analysis, then such an analysis clearly reduces the heterogeneity – only in business students' judgment achievements are moderator variables indicated. Surprisingly, students in business science clearly reach better judgment achievement than experts. This conclusion is also in line with our idiographic analysis. Moreover, experts and students in psychology science have a low judgment achievement, although, only students reach a high knowledge level. Hence, besides the low knowledge level, there have to be other factors responsible for the low judgment achievement in psychological science. Our results should be taken with caution, however, because the sample in each category is small, which leads also to publication bias indications (see below).

Finally, we want to add, that perhaps the *number of used cues* influences judgment achievement. As can be seen in the scatter plots as well as forest plots; each of them includes studies ordered according to the number of cues within the experience level in a research area. Our plots support the hypotheses that not only the number of cues and the research area should be considered, but also the types of cues – if we get a quantitative type of cue, for example a temperature measurement (see Stewart et al., 1997), or simply a description of the cue, for example a video (see Bernieri et al., 1996; Reynolds & Gifford, 2001). It is also subject to critique in our study that the cue-intercorrelation is not considered because of missing data. Therefore, in the future, we recommend that studies using the *LME* should also report the intercorrelation of the cues they use (see also Cooksey, 1996, p. 318). Finally, we want to refer to Kaufmann and Athanasou (2009), where a detailed illustration of the extreme points of correlation in association with the number of cues used in judgment tasks is presented. However, as it is the case with the reported data base – the number of judgment tasks in each category using only one cue leads to samples too small to answer this question satisfactorily. However, at first glance, there is no tendency revealed that fewer used cues – as suggested by the Fast and Frugal Heuristic Approach (Gigerenzer et al., 1999) – clearly lead to better judgment achievement. Further research is needed, however. In addition, we want to mention that with the used *LME* representing the cues in an additive way implying that judgment achievement increases with the increase of the number of cues. Hence, maybe the equation does not represent the environment or the decision maker's policy with absolute accuracy. However, before answering this question satisfactorily (whether the number of cues systematically influence judgment achievement) we have to consider the mentioned intercorrelation between the cues, the type of cues and finally, also the aggregation level of the cues.

As a highlight, not only the intercorrelation between the cues should be considered in the interpretation of our results, but also the introduced *intercorrelation* of the *LME* components. The high positive correlation across research areas between the *LME* components, except the *C* component, is not confirmed, if we focus on the different research areas and experience levels. Hence, the low judgment-achievement value found in psychological research – especially of psychological experts – could be explained by the negative intercorrelation between judgment achievement with task predictability and with consistency. This means that an increase in task predictability and in consistency would lead to a decrease of judgment achievement. It must be mentioned that such a negative intercorrelation between the *LME* components is not found in psychological students' *LME* components, although they reach a low judgment-achievement level too. Hence, this negative intercorrelation of *LME* components is rather associated with the *G* components, as students get a higher value compared to experts.

To summarize our meta-analysis, the main interest of this work is to clarify whether judgment achievement is stable across different research areas, as suggested by the Brunswikian tradition. We conclude that such an analysis is in line with the results found by Karelaia and Hogarth (2008) of an overall moderate judgment achievement. However, we supplement this conclusion as to that judgment achievement over all research areas is clearly not stable and in judgment achievement, the analysis should be done first from an individual perspective, then including studies from the same research areas – separated by the experience level – and finally, aggregated across judgment tasks considering also the aggregation principle as suggested by Wittmann (1985). From our point of view, only such a comprehensive analysis would reveal the importance of the underlying heterogeneity found in both data bases and consider possible aggregation bias. In addition, only such a procedure can in further research answer the question satisfactorily (whether the number of cues used in judgment tasks actually influences judgment achievement). The observed results support the ability of human estimating and decision-making, except in the area of psychology. This could be explained by low *G* values and negative *LME*-component intercorrelation. In addition, experts' judgment achievement is clearly not better than students' judgment achievement – across and within areas. However, students' business judgment achievement still indicates some moderator variables. As in this category two studies have no cue information, our hypothesis that the number of cues also influences judgment achievement can't be checked, hence, this questions is still open.

Before we focus on the limitations of our studies, we would like to mention our different robustness analyses to check our introduced results.

First, our *publication-bias calculations*. Although it is recommended by meta-analytic research to check the robustness of the results by means of publication-bias calculations, we were surprised by the resulting heterogeneity of our publication bias estimations. First, our publication analysis with Owen's fail-safe number of 61 indicated no publication bias across judgment tasks at first glance. In addition, our publication-bias analysis supplemented by the trim-and-fill method does not completely support the robustness of our results. However, we would like to mention that the underlying data base was heterogeneous, also leading to problems in such an analysis. Hence, Rothstein (2008, p. 78) concludes: "…be cautioned, that the effects of publication bias can be hard to disentangle from other sources of heterogeneity…", and we also consider this in our data interpretation, as our data base is heterogeneous as well. In addition, in association with our publication-bias estimations we would like to mention that this calculation is normally only applied to bare-bones meta-analyses. We are not aware of any psychometric meta-analysis using publication-bias analysis supplemented by the trim-and-fill method and would also like to highlight the novelty of the research areas on publication bias leading to appropriate estimates in the future and recommend to take our analysis with caution, especially the results in business and educational science.

Secondly, as introduced, we use a random-effect model, although most meta-analysts use fixed-effects models. However, our sensitivity analysis represents that such a model leads to a conservative estimation.

Thirdly, in our robustness analysis we checked the different weighting strategies used. In our idiographic-data based meta-analysis we used the number of profiles used in every task as a weighting factor. On the other hand, in our nomothetic-data based meta-analysis we used the number of judges as a weighting factor. After our robustness analysis we conclude that the weighting strategy used with our idiographic-data based meta-analysis leads to more indications of moderator factors in comparison to the weighting strategy used with our nomothetic data base. However, as the estimated average judgment achievement is the same, we use the strategy also recommended by Hunter and Schmidt (2004) and weight only the number of judges. Moreover, a profile-weighting strategy leads also to an exclusion of studies in which the number of judges' profiles is unknown – such as was the case with the study by Reynolds and Gifford (2001). Hence, such a strategy would limit the generalization of our results.

Fourthly, we assume that the study characteristic type of correlation (pearson vs. non-pearson correlation) does not consistently influence the *LME* components, as shown in our previous analysis (see Kaufmann & Athanasou, 2009).

To summarize, our meta-analysis based on the nomothetic data base has to be seen as a rather conservative estimation by our robustness analysis, first in relation to other reviews in the field, as we used a random-effect model. Secondly, also in comparison to our idiographic-data based analysis, in which we used another weighting strategy. Finally, as introduced in our discussion on the idiographic-data based analysis, also in the nomothetic data base analysis our artefact correction is rather conservative, as we did not correct any objective criterions.

So, if we compare the idiographic and the nomothetic data base, first from a bare-bones meta-analysis point of view – it's visible that not the same pattern is found of differences in research areas. Hence, we want to mention that in our idiographic-based data only 30 different tasks representing only a small sample of judgment tasks are included, and therefore these results should be taken with caution. On the other hand, the introduced great variation found in our idiographic database is visible in our scatter plots and confirmed by our conservative estimation based on the nomothetic data base.

In this study, every effort was undertaken to conduct an extensive analysis of the literature and to obtain reliable and valid findings that would aid enhancement and enrichment of judgment achievement in the framework of *SJT*. However, as with most research, there are some limitations to consider in the interpretation of the results found in this dissertation, which are presented next. The major points – *exclusion criteria, missing data, diagnostic vs. prognostic tasks, lack of task independency, LME critiques*, and *vicarious functioning concept* – are discussed in the following. More information is taken from Kaufmann and Athanasou (2009). In addition, the following critique points are supplemented by research recommendations which overcome the presented limitations of our study.

First, our *exclusion criteria* can be criticised, because we only considered achievement studies – feedback and learning studies were excluded. We see this selection as an advantage, as this represents daily judgments. Normally, we seldom get any feedback, which is the basis for learning. For example, in a hospital, physicians rarely get feedback about their decisions (see Katsikopoulos et al., 2008). Hence, our studies also represent realistic situations than is perhaps the case with feedback or learning studies. However, in a further analysis, the control group used in feedback and learning studies could be included and compared to increase the generalization of our results. Furthermore, a partial inclusion leading to a comparison would be interesting from the point of view, to exclude that any judgment achievement differences are already introduced by the study type as feedback, learning or achievement study. In addition, Karelaia's and Hogarth's work (2008) should be considered complementary to ours, as they completely included feedback as well as learning studies.

In addition, it should be mentioned that we excluded one study which used a dynamic situation (see Kirlik, 2006). This decision to only analyses studies which use a statistic situation was made to reduce possible variance caused by differences between dynamic and stable decision-making situations. However, if we compare these studies – with our judgment achievement values found in Figure 9 it becomes clear that this study reaches even higher judgment-achievement values, as in Kirlik's study the overall judgment achievement is .96, comparable to the values found in meteorology studies (Stewart, 1997). "Additionally, probability 95% of all the research in research in *JDM* is based on static tasks" (see Hammond, 2007, p. 238). Therefore, before an analysis like ours can conducted on dynamic situations, it is necessary to have more studies on these subjects to answer the research question, whether dynamic tasks are actually responsible for an increase in judgment achievement. This would be in line with Brunswik and call for representative design, in the meaning not to use questionnaires and change only single cues, leading to a new judgment situation which is quite unnatural of our daily life judgments. Generally, we suggest to transform the *LME* analysis from experimental studies to "naturalistic" studies, such as Kirliks' study (2006). In line with this suggestion, we want to add that most of the used *LME* studies in our meta-analysis are so-called *univariate Lens Models,* with exception of the study by Cooksey et al. (1986). However, the multivariable Lens Model is a better representation of a real-world decision maker.

Finally, we excluded studies aggregating their data across cues, such as Wittmann's (1985). In such an aggregation procedure, the individual variation is not eliminated. So, these studies imply an even greater individual heterogeneity than found in our studies. However, such studies would be ideal for inclusion in a further analysis and to focus on the differences between the aggregation procedures used in the studies (Wittmann, 1988). By how much would this type of aggregation study actually increase the heterogeneity in our data? Such an increase would even enforce individual differences in judgment achievement.

Secondly, a further limitation of our studies is that of *missing data* for a comprehensive analysis. Hence, we did not correct educational, psychological or other research areas for measurement errors with reported retest-reliability values, but used a theoretical estimation to check the robustness of the data. One can criticize our analysis in that the differences are introduced by our retest-reliability corrections. However, we want to emphasize that this is clearly not the case, because also our bare-bones meta-analysis already implies area differences. Finally, such a theoretical estimation of possible measurement errors is also advantageous in that our analysis is not bordered by the type of reliability values (e.g. consistency reliability, retest-reliability), as different reliability types also lead to different values. In addition, when regression models were estimated, bias-adjusted R^2 is well-established in linear regression models. Without such an adjustment, the correlation values could be overestimated. However, whether researchers used adjusted R^2 is not clearly visible from the studies, with some exceptions, such as Stewart (1990, 1997). Hence, we assume that the author used the adjusted R^2 calculations. Anyway, we calculated the bias-adjusted R^2 and rerun the overall meta-analysis (see Appendix I).

Thirdly, a further limitation of our analysis worth mentioning is that Wiggins (1973) differentiated *prognostics* and *diagnostics* tasks. We, however, did not focus on this aspect, but recommend including it in further analyses. Our hypothesis is that diagnosis reaches a better judgment achievement than prognosis, as this type of task also includes diagnoses (see Katsikopoulos et al., 2008), and a time constant has to be considered. Hence, this task type is more complex than diagnose tasks. Further research on this aspect is urgently needed to clarify the suggested hypothesis.

Fourthly, a further limitation in our research is that our analyses were affected by a *lack of independence* between correlations of judgment tasks in the same study (see Table 4). However, as this separation is often also made by the authors (e.g. Gorman et al., 1978; Stewart, 1997), we see this also in line with their research goal to find out the differences between judgment tasks. Had we not used such a strategy, it wouldn't have been possible to find out whether judgment achievement as in the study by Stewart (1997) is based on the number of cues or other task characteristics. Hence, we used the smallest unit of task to have as precise an analysis as possible to find the underlying judgment task characteristics maybe influencing the *LME* components.

Fifthly, our analysis *neglected LME critiques,* as it is often criticized the overfit of the linear regression used application. Overfitting is the case when too many informations are included in the models against the used number of informations by the judge. Hence, regression application also includes noise or simply too many free parameters – or irrelevant information for the judge. During the last decade, the Bayesian paradigm was used to control overfitting. This approach developed robust estimates of both environment and parameters, such as cues in our examples (for details, see Martignon & Hoffrage, 1999). Hence, a further Bayes theorem correction of our correlation values would be a proper estimation of our values (see Stewart, 1990). Consequently, in further analysis this fact should be included and added. If it is not possible to receive such values from literature, theoretical estimations could be used. Consequently, our results have to be interpreted with caution.

Sixthly, it must be mentioned that we only considered the achievement concept and neglected the underlying *vicarious-functioning concept,* because the *LME* does not detect the use of vicarious functioning by subject, nor does it measure the contribution of vicarious functioning to achievement. Hence, this work emphasizes that such research should urgently be done (for an example see Scholz & Tietje, 2002), because individuals adjust to the environment under so many different and changing circumstances.

Despite the mentioned limitations of our studies, our results achieve more transparency of judgment achievement and thus give researchers in this area a reference for their results. In addition, they get a better understanding of the sources of differences in *JDM* by means of the presentation of the components of the *LME* dependent on judgment tasks. This information also helps to identify task characteristics (e.g. number of cues) that may influence judgment achievement. Finally, our idiographic as well as nomothetic results enables the comparison of results with our meta-analysis in more detail. This is fruitful, as researchers in the psychological field can now get a comprehensive overview and realise the uniqueness of their results and follow our recommendation to focus on the *G* value in their research. Additional suggestions for further research are given in the following.

7 CONCLUSIONS AND DIRECTIONS FOR FUTURE RESEARCH

In the previous discussion we already introduced some research ideas. Besides these we emphasize in the following that further research should reveal why judgment achievement in meteorology studies is so high in comparison to psychological studies, implying research differences in judgment achievement.

First of all, we see the Hunter and Schmidt approach as a fruitful research tool in relation to classical statistical tests for further research on estimating judgment achievement. Hence, we would like to emphasize the following argumentations:

1) Compared to commonly used, classical statistical tests, the Hunter and Schmidt method uses random-effects instead of fixed-model models (see chapter 5.2.4.1).
2) Furthermore, although current statistic methods realise the importance of measurement error as longitudinal studies, none of them realise that there are other important artefacts, which should be considered as well (see chapter 4.4.4).
3) Finally, although the power of studies is discussed today (see Sedlmeier & Gigerenzer, 1989), Hunter and Schmidt already recognized that in meta-analysis research we have to be careful and look for other tools than significant tests (see chapter 4.5.1.1) to evaluate research (see Hunter & Schmidt, 2004, p. 8).

These three points clearly show Hunter and Schmidt's advanced thinking in statistics and their search for research tools to overcome the weaknesses of classical statistics. Furthermore, their permanent search for better research tools made this meta-analysis approach a useful tool for the future for both types of data base – idiographic and nomothetic.

The Hunter and Schmidt approach is still in development, and so this meta-analytic method has not yet reached an end stage of development. Hunter and Schmidt (2004) pointed out:

> All quantitative estimates are approximations. Even if these estimates are quite accurate, it is always desirable to make them more accurate, if possible. (p. 168)

Hence, our analysis could practically also support further correction in the direction of using the successful symmetry concept with the method of Hunter and Schmidt for a further artefact correction (Wittmann, 1988). As this concept is already being successfully applied in research, we see it as a supplement of the introduction to the Hunter and Schmidt meta-analytic method, so that further work will also be in line with Hunter and Schmidt's permanent intention to improve the approach. Hence, we could check whether high values found in meteorology in comparison to the low values in psychology are due to the asymmetry between the aggregated criterion and judgments (see Wittmann, 1985). In addition, according to Hammond (2007), we judge an objective criterion more accurately than a subjective one. However, our results that psychological judgments – mostly evaluated against subjective criteria are less accurate than meteorological judgments-evaluated against objective criteria support this hypothesis at first glance. Also education science used subjective criteria for their evaluation, but reaches a high judgment achievement level. Hence, this hypothesis clearly needs more research.

Beside this, many further questions arise, such as: Do the components of correlations of the *LME* vary systematically with demographical data of the judges (e.g. gender, age)? Or can the persons included in this meta-analysis be categorized according to their components of the *LME*? Do special judgment or task types exist? Furthermore, the studies could also be described in relation to the introduced *CCT* (see Hammond, 2000) to emphasise the value of the extern validity in *LME* studies – however, the internal validity is totally ignored in our analysis. Therefore, in a further meta-analysis, studies from Box 2, 3 (see Figure 2) should be included and compared to each other.

As introduced, with the *LME* component it is also possible to calculate the success *of expert models*. According to Wiggins (1973), this data base for expert modeling is his first suggested rule of thumbs for expert modeling: When criterion information exists, collect it and use it to construct statistical models of data combination (p. 220). Hence, such an analysis is highly recommended. The backgrounds for expert models are that they are superior to human judgment (Camerer, 1981; Grove et al., 2000), except in a “broken leg cue” situation. A “broken leg cue” situation resembles a situation in which a condition is changed so quickly that a constructed model can't react to it. But human beings are capable of adapting to the new environmental situation quite fast. In addition, Karelaia and Hogarth (2008) concluded that: High heterogeneity … further highlights the importance of identifying the task and judge characteristics that favour bootstrapping (p. 419). In addition, our analysis reveals also high variability underlying all *LME* components, so the question arises, whether a pattern could be found in the way that there are tasks in which expert models are useful or not. Especially in psychological science (or prediction of violence, see also Aegisdottir et al., 2006, p. 368) the success of expert models would be useful to overcome the low judgment-achievement level found in our study. Therefore, we are looking forward to the first meta-analysis according to Hunter-Schmidt (2004), considering in more detail expert models in the same way as the already done analysis based on idiographic and nomothetic data (see Appendix J, Kaufmann, Sjödahl, Athanasou & Wittmann, 2009). The nomothetic analysis could also directly be comparable with the study by Armstrong (2001) implying area differences in expert models applications.

However, an alternative to cognitive modelling to the regression approach is the Fast and Frugal Heuristic Approach. There are numerous articles comparing both approaches and leading to the superiority of the Fast and Frugal Heuristic Approach (Gigerenzer et al., 1999; Smith & Gilhooly, 2006). On the other hand, there are several critics referring to the fact that we neglected the assumption for a regression analysis in the Fast and Frugal Heuristic Approach, therefore this approach is superior. However, we showed that with our conservative artefact-corrected estimation that accuracy clearly increases. Hence, our analysis perhaps supports the superiority of the regression approach in comparison with the Fast and Frugal Heuristic Approach in some judgment tasks. However, there are no studies available at the moment comparing directly artefact-corrected judgment accuracy with the judgment achievement in the *LME* approach with the Fast and Frugal Heuristic Approach using an array of judgment tasks – ideally such research should be done with the suggested research design by Wittmann (1985) based on the Catell-boxes, using several single persons, tasks, and criterions (see Wittmann & Klumb, 2004). We would like to highlight that the use of real tasks – the ambulatory assessment approach (Fahrenberg, 2006) – could be an ideal approach for such research. This research would be a 2000 version of Brunswik's research on perception constancy in Berkley.

With the same type of research, also critique to the Lens Model research can be overcome. For example, Kirlik (2006) criticised the lens model, and therefore also the *LME* research as follows:

> One deficiency of the traditional lens model is that it portrays a view of the organism without any control over the environmental structure to which it must adapt. This is because there are no resources within that model to describe how an organism might use action to adapt the environment given its own needs and capacities for actions. (p. 214)

In line with this critique further research should also consider the aspect of action by including judgment achievement studies. Hence, not only is the adaptation to the environment also the adaptation of the environment, such as actions necessary to enhance our knowledge about internal cognition. For example, if you look around where you are at the moment, you will perhaps see some post-it messages like, tomorrow library, deadline of submission, today valentine's day. As you can see, you also adapt the environment, in that you would forget all these things if you haven't written them down.

In summary, although our study reveals some limitations and leads to suggestions for further research, the motivation for this work, which could also inspire further researcher is based on the following statement by Hammond (1996):

> I came to the conclusion after his [Brunswik] suicide that the best strategy was not to present and argue for the entire Brunswik approach, but to carry out empirical research on specific topics that at least some psychologists (and graduate students) would find interesting. In short, small deeds would have to speak louder than provocative words. So I took pieces of Brunswikian theory and method and went to work with these, … (p. 245)

In line with this statement, we hope that the presented work is one step in the direction to critically evaluate Brunswik’s suggested theory and method with empirical facts and to reinspire an academic discussion about Brunswik’s work and his value for the future of psychology research, mostly for a fruitful methodological approach in cognitive psychology.

REFERENCES

References marked with an asterisk (*) indicate studies included in the meta-analysis.

Aegisdottir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nicholos, C. N., Lampropoulos, G. K., Walker, B. S., Cohen, G., & Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist, 34*(3), 341-382.

Aguinis, H., Sturman, M. C. & Pierce, C. A. (2008). Comparison of three meta-analytic procedures for estimating moderating effects of categorical variables. *Organizational Research Methods, 11*(1), 9-34.

Alker, H. S. (1969). A typology of ecological fallacies. In M. Dogan & S. Rokan (Eds.), *Quantitative Ecological Analysis in the Social Sciences* (p. 69-86). Massachusetts: MIT Press.

Allport, G. W. (1937). *Personality: A psychological interpretation*. New York: Holt.

American Psychological Association (APA). (1954). Technical recommendations for psychological tests and diagnostic techniques. [Supplement]. *Psychological Bulletin, 51(2),* 1-38.

Anderson, N. H. (1981). *Foundations of information integration theory.* New York: Academic Press.

Armstrong, J. S. (2001). Judgmental bootstrapping: Inferring experts' rules for forecasting. In J. S. Armstrong (Ed.), *Pinciples of forecasting* (pp. 171-192). Philadelphia, Pennsylvania, USA: Springer.

Asendorpf, J. B. (2000). Idiographische und nomothetische Ansätze in der Psychologie [Idiographic and nomothetic approaches in psychology]. *Zeitschrift für Psychologie / Journal of Psychology, 208*, 72-90.

*Ashton, A. H. (1982). An empirical study of budget-related predictions of corporate executives. *Journal of Accounting Research, 20*(2), 440-449.

Ashton, R. H. (2000). A review and analysis of research on the test-retest reliability of professional judgment. *Journal of Behavioral Decision Making, 13*(3), 277-294.

Athanasou, J. A., & Cooksey, R. W. (1993). Self-estimates of vocational Interests. *Australian Psychologist, 28*(2), 118-1127.

*Athanasou, J. A., & Cooksey, R. W. (2001). Judgment of factors influencing interest: An Australian study. *Journal of Vocational Education Research, 26*(1), 1-13.

Balzer, W. K., Doherty, M. E., & O'Connor, R. (1989). Effects of cognitive feedback on performance. *Psychological Bulletin, 106*(3), 410-433.

Bangert-Drowns, R. L. (1986). Review of developments in meta-analytic method. *Psychological Bulletin, 99*(3), 388-399.

Baron, J. (2004). Normative models of judgment and decision making. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 19-36). Oxford: Blackwell Publishing.

Beach, L. R. (Ed.). (1990). *Image theory: Decision making in personal and organizational contexts*. Chichester, New York: John Wiley & Sons.

*Bernieri, F. J., Gillis, J. S., Davis, J. M., & Grahe, J. E. (1996). Dyad rapport and the accuracy of its judgment across situations: A lens model analysis. *Journal of Personality and Social Psychology, 71*(1), 110-129.

Bisantz, A. M., & Pritchett, A. R. (2003). Measuring the fit between human judgments and automated alerting algorithms: A study of collision detection. *Human Factors, 45,* 266-280.

Brehmer, B. (1976). Note on clinical judgment and the formal characteristics of clinical task. *Psychological Bulletin, 83*(5), 778-782.

Brehmer, B. (1987). Social judgment theory and forecasting. In G. Wright & P. Ayton (Eds.), *Judgmental forecasting* (pp. 199-216). Chichester: John Wiley & Sons.

Brehmer, B. (1988). The development of social judgment theory. In B. Brehmer & C. R. B. Joyce (Eds.), *Human judgment: The SJT view* (pp. 13-40). Amsterdam: North Holland.

Brehmer, A., & Brehmer, B. (1988). What have we learned about human judgment from thirty years of policy capturing. In B. Brehmer & C. R. B. Joyce (Eds.), *Human judgment: The SJT view* (pp. 75-114). Amsterdam: North Holland.

Brunswik, E. (1939). Probability as a determiner of rat behavoir. *Journal of Experimental Psychology, 25*, 175-197.

Brunswik, E. (1943). Organismic achievement and environmental probability. *Psychological Review, 50,* 255-272.

Brunswik, E. (1944). Distal focussing of perception: Size constancy in a representative sample of situations. *Psychological Monographs, 56*(254), 1-49.

Brunswik, E. (1952). The conceptual framework of psychology. *International encyclopedia of unified science.* Chicago, IL: University of Chicago Press.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62*(3), 193-217.

Brunswik, E. (Ed.). (1956). *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley, CA: University of California Press.

Brunswik, E. (1957). Scope and aspects of the cognitive problem. In J. S. Bruner, E. Brunswik, L. Festinger, F. Heider, K. F. Muenzinger, C. E. Osgood & D. Rapaport (Eds.), *Contemporary approaches to cognition* (p. 5-31). Cambridge: Harvard University Press.

Brunswik, E. (1966). Reasoning as a universal behavior model and a functional differentiation between "perception" and "thinking". In K. R. Hammond (Ed.), *The psychology of Egon Brunswik* (pp. 487-494). New York: Holt, Rinehart and Winston.

Camerer, C. (1981). General conditions for the success of bootstrapping models. *Organizational Behavior and Human Performance, 27*(3), 411-422.

Campbell Collaboration (2009). *Campbell Collaboration Guidelines*. Retrieved January 7, 2009, from http://campellcollaboration.org/resources/guidelines.php

Castellan, N. J. (1972). The analysis of multiple criteria in multiple-cue judgment tasks. *Organizational Behavior and Human Performance, 8*(2), 242-261.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Cooksey, R. W. (1996). *Judgment analysis: Theory, methods, and applications.* San Diego, CA: Academic press.

Cooksey, R. W., & Freebody, P. (1985). Generalized multivariate lens model analysis for complex human inference tasks. *Organizational Behavior and Human Decision Processes, 35*, 46-72.

Cooksey, R. W., Freebody, P. & Bennett, A. J. (1990). The ecology of spelling: A lens model analysis of spelling errors and student judgments of spelling difficulty. *Reading Psychology: An International Quarterly, 11,* 293-322.

*Cooksey, R. W., Freebody, P., & Davidson, G. R. (1986). Teachers' predictions of children's early reading achievement: An application of social judgment theory. *American Educational Research Journal, 23*(1)*,* 41-64.

Cooksey, R. W., Freebody, P., & Wyatt-Smith, C. (2007). Assessment as judgment-in-context: Analysing how teachers evaluate students' writing. *Educational Research and Evaluation, 13,* 401-434.

Cooper, H., & Hedges, V. L. (Eds.). (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.

*Cooper, R. P., & Werner, P. D. (1990). Predicting violence in newly admitted inmates: A lens model analysis of staff decision making. *Criminal Justice and Behavior, 17*(4), 431-447.

Dalgleish, L. I. (1988). Decision making in child abuse cases: Application of SJT and signal detection theory. In B. Brehmer & C. R. B. Joyce

(Eds.), *Human judgment: The SJT view* (pp. 317-360). Amsterdam: North Holland.

Dawes, R. M. (1971). Case study of graduate admission: Application of three principles of human decision making. *American Psychologist, 26*(1), 180-188.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*(7), 571-582.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*(2), 95-106.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials, 7,* 177-188.

Dhami, M. K., Hertwig, R., & Hoffrage, R. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin, 130*(6), 959-988.

Dodrill, C. B. (1983). Long-term reliability of the Wonderlic personnel test. *Journal of Consulting and Clinical Psychology, 51*(2), 316-317.

Dougherty, T. W., Ebert, R. J., & Callender, J. C. (1986). Policy capturing in the employment interview. *Journal of Applied Psychology, 71,* 9-15.

Dunwoody, P. T. (2006). The neglect of the environment by cognitive psychology. *Journal of Theoretical and Philosophical Psychology, 26*, 139-153.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56,* 455-463.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment* (pp. 17-52). New York: Wiley.

Edwards, W. & Newman, J. R. (1982). *Multiattribute evaluation*. Beverly Hills. CA: Sage.

Egger, M., Smith, G. D., & Altman, D. G (2001). *Systematic reviews in health care: Meta-analysis in context.* London, UK: British Medical Journal Publication Group.

*Einhorn, H. J. (1974). Cue definition and residual judgment. *Organizational Behavior and Human Performance, 12*(1), 30-49.

Eysenck, H. J. (1952). The effects of psychotherapy: An evaluation. *Journal of Consulting Psychology, 16,* 319-324.

Eysenck, H. J. (1994). Systematic reviews: Meta-analysis and its problems. *British Medical Journal, 309*, 789-792.

Fahrenberg, J. (2006). *Assessment in daily life. A review of computer-assisted methodologies and applications in psychology and psychophysiology, years 2000-2005.* Retrieved October 4, 2008, from http://www.ambulatory-assessment.org/

Field, A. P. (2001). Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed- and random-effects methods. *Psychological Methods, 6*(2), 161-180.

Field, A. P. (2005). Is the meta-analysis of correlations accurate when population correlations vary? *Psychological Methods, 10*(4), 444-467.

Gigerenzer, G. (2007). *Gut feelings. The intelligence of the unconscious*. New York: Viking Press.

Gigerenzer, G., Todd, P. M., & the ABC Research Group (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*, 3-8.

Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological Bulletin, 73*, 422-432.

*Goldberg, L. R. (1976). Man versus model of man: Just how conflicting is that evidence? *Organizational Behavior and Human Performance, 16*(1), 13-22.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review, 109*(1), 75-90.

Goldstein, W. M. (2004). Social judgment theory: Applying and extending Brunswik's probabilistic functionalism. In D. J. Koehler & N. Harvey

(Eds.), *Blackwell handbook of judgment and decision making* (pp. 37-67). Oxford: Blackwell Publishing.

*Gorman, C. D., Clover, W. H., & Doherty, M. E. (1978). Can we learn anything about interviewing real people from "interviews" of paper people? Two studies of the external validity of a paradigm. *Organizational Behavior and Human Performance, 22*(2)*,* 165-192.

Grebstein, L. C. (1963). Relative accuracy of actuarial prediction, experienced clinicians and graduate students in a clinical judgment task. *Journal of Consulting Psychology, 27,* 127-132.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*(1), 19-30.

Hall, S. M., & Brannick, M. T. (2002). Comparison of two random-effects methods of meta-analysis. *Journal of Applied Psychology, 87*(2), 377-389.

Hammond, K. R. (1955). Probabilistic functioning and the clinical method. *Psychological Review, 62*(4), 255-262.

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

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

Hammond, K. R. (1998). *Ecological validity: Then and now*. Retrieved April 3, 2006, from http://www.brunwik.org/notes/essay2.html

Hammond, K. R. (2000). *Judgments under stress*. New York: Oxford University Press.

Hammond, K. R. (Ed.). (2007). *Beyond rationality: The search for wisdom in a troubled time*. New York: Oxford University Press.

Hammond, K. R., Hamm, R. M., & Grassia, J. (1986). Generalizing over conditions by combining the multitrait-multimethod matrix and the representative design of experiments. *Psychological Bulletin, 100*(2), 257-269.

Hammond, K. R., Hursch, C. J., & Todd, F. J. (1964). Analyzing the components of clinical inference. *Psychological Review, 71*(6), 438-456.

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

Hammond, K. R., Stewart, T. R., Brehmer, B., & Steinmann, D. O. (1975). Social judgment theory. In M. F. Kaplan & S. Schwartz (Eds.), *Human judgment and decision processes* (pp. 271-317). New York: Academic Press, Inc.

Harvey, N. (1995). Why are judgments less consistent in less predictable task situations? *Organizational Behavior and Human Decision Processes, 63*(3), 247-263.

*Harvey, N., & Harries, C. (2004). Effects of judges' forecasting on their later combination for forecasts for the same outcomes. *International Journal of Forecasting, 20*(3)*,* 391-409.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

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

Hirst, M. K., & Luckett, P. F. (1992). The relative effectiveness of different types of feedback in performance evaluation. *Behavioral Research in Accounting, 4,* 1-22.

Holzworth, J. (1999). *An annotated bibliography of all published cue probability learning studies.* Retrieved March 21, 2005, from http://www.brunswik.org/resources/mcplbib.doc

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies.* Beverly Hills, CA: Sage Publications.

Hurlburt, R. T., & Knapp, T. J. (2006). Münsterberg in 1898, not Allport in 1937, introduced the terms idiographic and nomothetic to American psychology. *Theory & Psychology, 16*(2), 287-293.

Hursch, C. J., Hammond, K. R., & Hursch, J. L. (1964). Some methodological considerations in multiple-cue probability learning studies. *Psychological Review, 71*(1), 42-60.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases.* London: Cambridge University Press.

Karelaia, N., & Hogarth, R. (2008). Determinants of linear judgment: A meta-analysis of lens studies. *Psychological Bulletin, 134*(3), 404-426.

Katsikopoulos, K. V., Pachur, T., Machery, E., & Wallin, A. (2008). From Meehl (1954) to fast and frugal heuristics (and back): New insights into how to bridge the clinical - actuarial divide. *Theory & Psychology, 18*(4), 443-463.

Kaufmann, E. (2006, Nov.). Judgment achievement under the lens. *The Brunswik Society Newsletter, 21,* 10.

Kaufmann, E. (2007, Nov.). Call for more idiographic-nomothetic based judgment achievement research. *The Brunswik Society Newsletter, 22,* 12-13.

Kaufmann, E., & Athanasou, J. A. (2009). A meta-analysis of judgment achievement defined by the lens model equation. *Swiss Journal of Psychology, 68* (2), 99-112.

Kaufmann, E., & Sjödahl, L. (2006, Nov.). The idiographic approach in social judgment theory: A meta-analysis of components of the lens model equation. *The Brunswik Society Newsletter, 21,* 11-12.

Kaufmann, E., Sjödahl, L., Athanasou, J. A., & Wittmann, W. W. (2007, Nov.). A critical meta-analytic perspective of the components of the lens model equation in judgment achievement. *The Brunswik Society Newsletter, 22,* 14-15.

Kaufmann, E., Sjödahl, L., Athanasou, J. A., & Wittmann, W. W. (2008, Nov.). On Brunswik's trace in achievement studies. *The Brunswik Society Newsletter, 23,* 24.

Kaufmann, E., Sjödahl, L., Athanasou, J. A., & Wittmann, W. W. (2009). Do we underestimate the validity of expert models? [Working paper]. Mannheim, Germany: University of Mannheim.

Kaufmann, E., Sjödahl, L., & Mutz, R. (2007). The idiographic approach in social judgment theory: A review of components of the lens model equation components. *International Journal of Idiographic Science, 2.*

*Kim, C. N., Chung, H. M., & Paradice, D. B. (1997). Inductive modeling of expert decision making in loan evaluation: A decision strategy perspective. *Decision Support Systems, 21*(2), 83-98.

Kirlik, A. (2006). *Human-technology interaction: Methods and models for cognitive engineering and human-computer interaction*. Oxford: University Press.

Kisamore, J. L., & Brannick, M. T. (2008). An illustration of the consequences of meta-analysis model choice. *Organizational Research Methods, 11*(1), 35-53.

Klein, G. A., Orasanu, J., Calderwood, R., & Zasambok, C. E. (Eds.). (1993). *Decision making in action: Models and methods*. New Jersey: Ablex Publishing.

Kreppner, K. (1992). William L. Stern, 1871-1938: A neglected founder of developmental psychology. *Developmental Psychology, 28*(4), 539-547.

*LaDuca, A., Engel, J. D., & Chovan, J. D. (1988). An exploratory study of physicians' clinical judgment: An application of social judgment theory. *Evaluation & the Health Professions, 11*(2), 178-200.

Lamiell, J. T. (2003). *Beyond individual and group differences: Human individuality, scientific psychology, and William Stern's critical personalism.* Thousand Oaks, CA: Sage Publications.

Lee, J. W., & Yates, J. F. (1992). How quantity judgment changes as the number of cues increases: An analytical framework and review. *Psychological Bulletin, 112*(2), 363-377.

*Lehman, H. A. (1992). *The prediction of violence by lay persons: Decision making by former psychiatric inpatients.* Unpublished doctoral dissertation, The California School of Professional Psychology Berkeley/Alameda.

*Levi, K. (1989). Expert systems should be more accurate than human experts: Evaluation procedures from human judgment and decision making. *IEEE Transactions on Systems, Man, and Cybernetics, 19*(3), 647-657.

Light, R. J., & Pillemenr, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48*(12), 1181-1209.

Lyons, K. D., Tickle-Degnen, L., Henry, A., & Cohn, E. (2004). Impressions of personality in parkinson's disease: Can rehabilitation practitioners see beyond the symptoms? *Rehabilitation Psychology, 49*(4), 328-333.

*MacGregor, D., & Slovic, P. (1986). Graphic representation of judgmental information. *Human-Computer Interaction, 2,* 179-200.

Martignon, L., & Hoffrage, U. (1999). Why does one-reason decision making work? A case study in ecological rationality. In: Gigerenzer, G. Todd, P. M. & the ABC Research Group (eds.). *Simple heuristics that make us smart* (pp. 119-140). Oxford University Press.

*McClellan, P. G., Bernstein, I. H., & Garbin, C. P. (1984). What makes the Mueller a liar: A multiple-cue approach. *Perception & Psychophysics, 36*(3), 234-244.

*Mear, R., & Firth, M. (1987). Assessing the accuracy of financial analyst security return predictions. *Accounting Organizations and Society, 12*(4), 331-340.

Meehl, P. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN: University of Minnesota Press.

Miller, G. (1956). The magical number seven plus or minus two: Some limits to our capacity for processing information. *The Psychological Review, 63*(2), 81-97.

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement, 2*(4), 201-218.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Nunnally, J., & Bernstein, I. (1994). *Psychometric theory*. New York: McGraw-Hill.

Nystedt, L. (1974). Consensus among judges as a function of amount of information. *Journal of Educational and Psychological Measurement, 34*(1), 91-101.

Nystedt, L., & Magnusson, D. (1972). Predictive efficiency as a function of amount of information. *Multivariate Behavioral Research, 7*(4), 441-450.

*Nystedt, L., & Magnusson, D. (1975). Integration of information in a clinical judgment task, an empirical comparison of six models. *Perceptual and Motor Skills, 40*(2), 343-356.

O'Connor, M., Remus, W., & Lim, K. (2005). Improving judgmental forecasts with judgmental bootstrapping and task feedback support. *Journal of Behavioral Decision Making, 18,* 246-260.

Orwin, R. G. (1983). A fail-safe *N* for effect size in meta-analysis. *Journal of Educational Statistics, 8,* 157-159.

Over, D. (2004). Rationality and the normative / descriptive distinction. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 3-18). Oxford: Blackwell Publishing.

Pearson (1904). Report on certain enteric fever inoculation statistics. *British Medical Journal, 3*, 1243-1246.

R. (2007). R (Version 2.6.1) [Software]. The R foundation for Statistical Computation.

*Reynolds, d. A. J., & Gifford, R. (2001). The sounds and sights of intelligence: A lens model channel analysis. *Personality and Social Psychology Bulletin, 27*(2), 187-200.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15*(2), 351-357.

*Roose, J. E., & Doherty, M. E. (1976). Judgment theory applied to the selection of life insurance salesmen. *Organizational Behavior and Human Performance, 16*(2), 231-249.

Rosenthal, R. (1979). The "file drawer problem" and the tolerance for null results. *Psychological Bulletin, 86*(3), 638-641.

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. (Rev. ed). Newbury Park, CA: Sage.

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52*, 59-82.

Rothstein, H. R. (2008). Publication bias as a threat to the validity of meta-analytic results. *Journal of Experimental Criminology, 4*, 61-81.

Rubin, D. B. (1990). A new perspective on meta-analysis. In K. W. Wachter & M. L. Straf (Eds.), *The future of meta-analysis (pp. 155-165)*. New York: Russell Sage.

Runyan, W. M. (1983). Idiographic goals and methods in the study of lives. *Journal of Personality, 51*(3), 413-437.

Sackett, P. R., Harris, M. M., & Orr, J. M. (1986). On seeking moderator variables in meta-analysis of correlation data: A monte carlo investigation of statistical power and resistance to type I error. *Journal of Applied Psychology, 71*(2), 302-310.

Schmidt, F. L., Berner, J. G., & Hunter, J. E. (1973). Racial differences in validity of employment tests: Reality or illusion? *Journal of Applied Psychology, 58*(1), 5-9.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological reserach: Lessons from 26 research scenarios. *Psychological Methods, 1,* 199-223.

Schmidt, F. L., & Le, H. A. (2005). Hunter-Schmidt meta-analysis programs (Version 1.1) [Computer software]. University of Iowa, Department of Management & Organization, Iowa City, IA 42242.

Scholz, R. W., Mieg, H. A., & Weber, O. (2003). Wirtschaftliche und organisationale Entscheidungen [Business and organisational decisions]. In A. E. Auhagen & H. W. Bierhoff (Eds.), *Wirtschafts- und Organisationspsychologie* (pp. 194-219). Weinheim: Beltz.

Scholz, R. W., & Tietje, O. (2002). *Embedded case study methods: Integrating quantitative and qualitative knowledge*. Thousand Oaks: Sage.

Schulze, R. (2007). The state and the art of meta-analysis. *Zeitschrift für Psychologie / Journal of Psychology, 215*(2), 87-89.

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

Shadish, W. R. (2007). *William R. Shadish*. Retrieved January 7, 2009, from http://faculty.ucmerced.edu/wshadish/index.htm

Shanteau, J. (2001). Management decision making. In W. E. Craighead & C. B. Nemeroff (Eds.), *Encyclopedia of psychology and behavioral science* (pp. 913-915). NY: Wiley.

Shanteau, J. (2002). *Domain differences in expertise.* Working paper. Kansas State University, KS: Manhattan.

Shanteau, J., & Stewart, T. R. (1992). Why study expert decision making? Some historical perspectives and comments. *Organizational Behavior and Human Decision Processes, 53*(2), 95-106.

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics, 69,* 99-118.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review, 63*(2), 129-138.

*Singh, H. (1990). Relative evaluation of subjective and objective measures of expectations formation. *Quarterly Review of Economics and Business, 30*(1), 64-74.

Slavin, R. E. (1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational Researcher, 15,* 5-11.

Smith, C. T. & Williamson, P. R. (2007). A comparison of methods for fixed effects meta-analysis of individual patient data with time to event outcomes. *Clinical Trials, 4*(6), 621-630.

Smith, L., & Gilhooly, K. (2006). Regression versus fast and frugal models of decision-making: The case of prescribing for depression. *Applied Cognitive Psychology, 20*(2), 265-274.

*Smith, L., Gilhooly, K., & Walker, A. (2003). Factors influencing prescribing decisions in the treatment of depression: A social judgment theory approach. *Applied Cognitive Psychology, 17*(1)*,* 51-63.

Smith, M. L. & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32,* 752-760.

*Speroff, T., Connors, A. F., & Dawson, N. V. (1989). Lens model analysis of hemodynamic status in the critically ill. *Medical Decision Making, 9*(4), 243-261.

SPSS, Inc. (2004). SPSS for Windows (Version 13.0). Chicago: SPSS, inc.

*Steinmann, D. O., & Doherty, M. E. (1972). A lens model analysis of a bookbag and poker chip experiment: A methodological note. *Organizational Behavior and Human Performance, 8*(3), 450-455.

Stewart, T. R. (1976). Components of correlations and extensions of the lens model equation. *Psychometrika, 41*(1), 101-120.

*Stewart, T. R. (1990). Notes and correspondence: A decomposition of the correlation coefficient and its use in analyzing forecasting skill. *American Meteorological Society, 5,* 661-666.

Stewart, T. R. (1997). Meta-analysis of the relation between task predictability and accuracy of expert judgment. *Center for Policy Research, University at Albany, State University of New York, Albany*.

Stewart, T. R. (2004, Nov.). Ledyard R. Tucker. *The Brunswik Society Newsletter, 19*, 2-3.

Stewart, T. R., & Lusk, C. M. (1994). Seven components of judgmental forecasting skill: Implications for research and the improvement of forecasts. *Journal of Forecasting, 13*(7), 579-599.

Stewart, T. R., Middleton, P., Downton, M., & Ely, D. (1984). Judgments of photographs versus field observations in studies of perception and judgment of the visual environment. *Journal of Environmental Psychology, 4,* 283-302.

Stewart, T. R., Moninger, W. R., Grassia, J., Brady, R. H., & Merrem, F. H. (1989). Analysis of expert judgment in a hail forecasting experiment. *Weather and Forecasting, 4*, 24-34.

*Stewart, T. R., Roebber, P. J., & Bosart, L. F. (1997). The importance of the task in analyzing expert judgment. *Organizational Behavior and Human Decision Processes, 69*(3), 205-219.

*Szucko, J. J., & Kleinmuntz, B. (1981). Statistical versus clinical lie detection. *American Psychologist, 36*(5), 488-496.

SYSTAT, Inc. (2000). SYSTAT for Windows (Version 10). [Computer software]. Evanston, IL: SYSTAT, Inc.

Tape, T. G., Heckerling, P. S., Ornato, J. P., & Wigton, R. S. (1991). Use of clinical judgment analysis to explain regional variations in physicians' accuracies in diagnosing pneumonia. *Medical Decision Making, 11*(3), 189-197.

Tickle-Degnen, L., & Lyons, K. D. (2004). Practitioners' impressions of patients with parkinson's disease: The social ecology of the expressive mask. *Social Science & Medicine, 58*(3), 603-614.

Todd, F. J. (1954). *A methodological analysis of clinical judgment.* Unpublished doctoral dissertation, University of Colorado.

Tolman, E. C., & Brunswik, E. (1935). The organism and the causal texture of the environment. *Psychological Review, 42*, 43-77.

*Trailer, J. W., & Morgan, J. F., (2004). Making "good" decisions: What intuitive physics reveals about the failure of intuition. *The Journal of American Academy of Business, 3*(1), 42-48.

Tucker, L. R. (1964). A suggested alternative formulation in the developments by Hursch, Hammond and Hursch and by Hammond, Hursch and Todd. *Psychological Review, 71*(6), 528-530.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131.

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.

Viechtbauer, W. (2007). Random-effects models and moderator analyses in meta-analysis. *Zeitschrift für Psychologie / Journal of Psychology, 215*(2), 104-121.

*Werner, P. D., Rose, T. L., Murdach, A. D., & Yesavage, J. A. (1989). Social workers’ decision making about the violent client. *Social Work Research & Abstracts, 25*(3), 17-20.

*Werner, P. D., Rose, T. L., & Yesavage, J. A. (1983). Reliability, accuracy, and decision-making strategy in clinical predictions of imminent dangerousness. *Journal of Consulting and Clinical Psychology, 51*(6), 815-825.

Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment*. Reading, MA: Addison-Wesley Publishing Company.

*Wiggins, N., & Kohen, E. S. (1971). Man versus model of man revisited: The forecasting of graduate school success. *Journal of Personality and Social Psychology, 19*(1), 100-106.

Windelband, W. (1894). Geschichte und Naturwissenschaft [History and natural science]. In *Praeludien* (Vol. 2, pp. 136-160). Tübingen.

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

Wittmann, W. W. (1988). Multivariate reliability theory. Principles of symmetry and successful validation strategies. In J. R. Nesselroade &

R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (pp. 505-560). New York: Plenum Press.

Wittmann, W. W., & Klumb, P. L. (2006). How to fool yourself with experiments in testing theories in psychological research. In R. R. Bootzin & P. E. McKnight (Eds.). *Strengthening research methodology: Psychological measurement and evaluation* (pp. 185-211). Washingtion D. C.: American Psychological Association.

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

Wood, J. A. (2008). Methodology for dealing with duplicate study effects in a meta-analysis. *Organizational Research Methods, 11*(1), 79-95.

Wolf, B. (1995). *Brunswick und ökologische Perspektiven in der Psychologie* [Brunswik and ecological perspectives in psychology]. Weinheim: Deutscher Studien Verlag (Habilitation).

Wright, D. (2005). *Meta-analysis of correlation coefficients (Schmidt-Hunter method)*. Retrieved January 7, 2006, from http://www.sussex.ac.uk/Users/danw/ESM/SHcorrmeta.SPS

*Wright, W. F. (1979). Properties of judgment models in a financial setting. *Organizational Behavior and Human Performance, 23*(1), 73-85.

APPENDICES

A

APPENDIX A: ABBREVIATIONS

C	Consistency component of the *LME*
CCT	Cognitive Continuum Theory
DL	DerSimonian and Laird estimator (1986)
FM	Fixed-effect models
G	Linear knowledge component of the *LME*
JDM	Judgment and Decision Making
LME	Lens Model Equation
nr	Study number according to Tables 5 and 6
r_0	Type of correlation is unknown
r_a	Judgment achievement
R_e	Environmental predictability component of the *LME*
R_s	Consistency component of the *LME*
RM	Random-effect models
rr	Retest-reliability value
SJT	Social Judgment Theory

APPENDIX B: LITERATURE SEARCH

B: Table 1

Results (hits and date) of our literature search in data bases

	Search engines				
	PsychArticles	PsycINFO	PSYNDEXplus	Eric	Eric Online
Keywords	hits/date	hits/date	hits/date	hits/date	hits/date
Social Judgment Theory	502/11.04.08	503/03.04.08	329/08.05.08	32/11.03.08	216/11.03.08
Social Judgement Theory	8185/11.04.08	507/03.04.08	2183/08.05.08	3/11.03.08	22/11.03.08
Lens Model Equation	540/11.04.08	269/03.04.08	56/08.05.08	2/11.03.08	2/11.03.08
Lens Model	551/11.04.08	608/07.05.08	882/08.05.08	7/11.03.08	46/11.03.08
Judgment achievement	530/03.04.08	2054/07.05.08	272/08.05.08	21/11.03.08	133/11.03.08
Judgement achievement	992/03.04.08	1263/07.05.08	442/08.05.08	11/11.03.08	30/11.03.08
Lens Model Analysis	503/03.04.08	802/07.05.08	390/08.05.08	0/11.03.08	355/11.03.08
Idiographic approach	89/03.04.08	221/07.05.08	69/08.05.08	0/11.03.08	48/11.03.08
Judgment accuracy	521/03.04.08	557/07.05.08	339/08.05.08	4/11.03.08	156/11.03.08
Judgement accuracy	1287/03.04.08	295/07.05.08	71/ 8.05.08	22/11.03.08	14/11.03.08

B: Table 2

Results (hits and date) of our literature search in (online) data bases

	Search engines				
	WebofScience	Google	Google.scholar	Yahoo.com	Social Science Research Network
Keywords	hits/date	hits/date	hits/date	hits/date	hits/date
Social Judgment Theory	46/12.03.08	1360/14.03.08	1480/07.08.08	16100/14.07.08	0/02.07.08
Social Judgement Theory	10/12.03.08	2440/06.08.08	3940/07.08.08	1480/02.07.08	0/02.07.08
Lens Model Equation	8/12.03.08	885/08.08.08	204/07.08.08	930/02.07.08	1/02.07.08
Lens Model	16/12.03.08	179000/14.03.08	6 680/07.08.08	1130/02.07.08	0/02.07.08
Judgment achievement	85/12.03.08	731/14.03.08	93/07.08.08	324000/02.07.08	3/02.07.08
Judgement achievement	53/12.03.08	192/14.03.08	12/07.08.08	66560/02.07.08	0/02.07.08
Lens Model Analysis	0/12.03.08	679/06.08.08	374/07.08.08	10900/02.07.08	0/02.07.08
Idiographic approach	30/12.03.08	6560/10.04.08	1930/07.08.08	175/02.07.08	0/02.07.08
Judgment accuracy	113/12.03.08	10900/06.08.08	1850/07.08.08	17300/02.07.08	2/02.07.08
Judgement accuracy	11/12.03.08	1730/06.08.08	379/07.08.08	1180/02.07.08	6/02.07.08

B: Table 3

Results (hits and date) of our literature search in German in the data base Wiso-Net

	Search engine
	Wiso-Net
Keywords	hits/date
Soziale Urteilstheorie	0/08.08.08
Linsen-Modell Gleichung	4/11.08.08
Linsen Model	28/11.08.08
Linsen Modell	224/11.08.08
Urteilsleistung	0/11.08.08
Linsen Modell Analyse	37/11.08.08
Idiographischer Ansatz	1/11.08.08
Urteilsgenauigkeit	2/11.08.08

APPENDIX C: *LME* COMPONENT CALCULATION

The *G* component in the *LME* (see Equation C: 1):

$$G = \frac{r_a - C\sqrt{1-R_e^2}\sqrt{1-R_s^2}}{R_s R_e} \qquad \text{(C: 1)}$$

The *C* component in the *LME* (see Equation C: 2)*:*

$$C = \frac{r_a - GR_sR_e}{\sqrt{1-R_s^2}\sqrt{1-R_e^2}} \qquad \text{(C: 2)}$$

The R_s component in the *LME* (see Equation C: 3):

$$R_s^{1/2} = \frac{GR_e r_a \pm \sqrt{G^2R_e^2r_a^2 - (G^2R_e^2 + C^2 - C^2R_e{}^2)(r_a^2 - C^2 + C^2R_e^2)}}{(G^2R_e^2 + C^2 - C^2R_e^2)} \qquad \text{(C: 3)}$$

APPENDIX D: COMPARISON WITH THE META-ANALYSIS BY KARELAIA AND HOGARTH (2008)

In the following Table D: 1, reasons for exclusion of studies in our meta-analysis are specified.

D: Table 1

Reasons for the exclusion of studies in our meta-analysis

Study	Reason for exclusion
Grebstein (1963) Todd (1954)	Study published before 1964
Brisantz & Pritchett (2003)	N-system lens model (see chapter 2.4.1)
Kirlik (2006)	Dynamic judgment task
Cooksey, Freebody, & Wyatt-Smith (2007)	Agreement between two policy capture models
Stewart, Middleton, Downton, & Ely (1984) Wittmann (1985)	Aggregation across cues
Cooksey, Freebody, & Bennett, 1990	Repeated tasks after one week
Dalgleish (1988) Hirst & Luckett (1992) O'Connor, Remus, & Lim (2005)	Feedback study
Doherty, Ebert, & Callender (1986)	Police capturing study (see chapter 2.4.1)

D: Table 2

A study list and the explanations for different coding in our data base in comparison to Karelaia and Hogarth (2008)

nr	Study	Explanation for the different coding in our data base:
12	Wright (1979)	In contrast to Karelaia and Hogarth, we didn't separate our studies into two groups of persons, as there are the same number of profiles, and the number of cues and also the component R_e are the same.
13	Harvey & Harries (2004)	This experiment showed that judges' ability to combine forecasts that they receive from more knowledgeable advisors is impaired when they have previously made their own forecasts for the same outcomes. We used only the baseline.
15	Cooksey, Freebody, & Davidson (1986)	As there are two criterions available, and relating to them the *LME* values, we coded these studies with two tasks, reading comprehension and word knowledge, instead of only one task as suggested by Karelaia and Hogarth (2008) (Univariate instead of multivariate Lens Model).
22	Gorman, Clover, & Doherty (1978)	As the authors described the lens-model components for the interview and the paper-people treatment and mention these as two experimental treatments, this represents two types of tasks for us. Also, the number of profiles varies.
27	Stewart, Roebber, & Bosart (1997)[a]	We separated this study into four tasks, as there are different numbers of cues, different numbers of profiles, as well as different time and weather forecasts. Each task also has different R_e values. Karelaia and Hogarth (2008) included them as one task.

Note. nr = study number according to Table 5 and 6.

[a]is coded as learning study by Karelaia and Hogarth (2008).

To summarize, five studies of the 19 overlapping studies are included with difference in separating in judgment task (see D: Table 1) leading to 14 studies. Hence, differences in the data-base of the remaining 14 studies are presented in the following. However, first, we will compare four study characteristics (see Table D: 2), then the *LME* components (see Tables D: 3, 4).

D: Table 3

Study-characteristics agreement with the data-base by Karelaia and Hogarth (2008)

nr	Study	Number of judges	Number of judgments	Number of cues	Expertise level
2	Levi (1989)	=	=	=	=
3	LaDuca et al. (1988)	=	=	=	=
4	Smith et al. (2003)	=	=	=	=
7	Ashton (1982)	=	=	=	=
8	Roose & Doherty (1976)	=	=	66(64/5)	=
11	Mear & Firth (1987)	=	=	12(10)	=
12	Wright (1979)	=	=	=	[a]
13	Harvey & Harries (2004)	=	=	[b]	=
15	Cooksey et al. (1986)	=	=	=	=
16	Wiggins & Kohen (1971)	=	90(110)[c]	=	=
17	Athanasou & Cooksey (2001)	=	=	=	=
18	Szucko & Kleinmutz (1981)	=	=	10(3, 4)	=
19	Cooper & Werner (1990)	10, 11 (18)[d]	=	=	=
20	Werner et al. (1989)	=	=	=	=
21	Werner et al. (1983)	=	=	=	=
22	Gorman et al. (1978)	=	57(75)	=	[e]
27	Stewart et al. (1997)	=	=	=	[f]
28	Steinman & Doherty (1972)	=	=	=	=
29	MacGregor & Slovic (1986)	=	=	=	=

Note. nr = study number according to Table 5 and 6. = data agreement, if the data does not agree, the Karelaia and Hogarth (2008) value is reported and supplemented by our value in parentheses.

[a] value can not be compared, because the study was separated into two groups by Karelaia and Hogarth (2008).

[b] it was not available.

[c] We used 110 profiles, like Armstrong (2001), in contrast to Karelaia and Hogarth (2008).

[d] Karelaia and Hogarth (2008) separated their data set in two groups (10 psychologists, 11 case managers). In our study, only the evaluation of nine psychologists and nine case managers were included, as footnotes mention that “one psychologist and two case managers consistently labelled every case as not violent. Consequently, these judges were dropped from within-judge correlation analyses involving predictive accuracy and components of the lens model” (Cooper & Werner, 1990, p. 445).

[e] Karelaia and Hogarth (2008) coded the experience level with training experience, hence, it is not directly comparable, as we didn't include such a category.

[f] We coded this study differently, separating students and experts, in contrast to Karelaia and Hogarth (2008), labelling all participants as experts.

To summarize: the 19 overlapping studies, showing a 92% agreement relating to study characteristics. However, 6 studies can’t be compared in relation to *LME* (see Table 2, plus the study by Cooper & Werner, 1990). Hence, in the following, the 13 studies are compared in relation to the *LME* components. In seven studies, or 50% of the studies, no differences relating to the *LME* components were found (see D: Table 4). The six studies with differences in *LME* components are reported in D: Table 5.

D: Table 4

The seven studies with no differences in the LME components

nr	Study
2	Levi (1989)
4	Smith (2003)
8	Roose & Doherty (1976)
11	Mear & Firth (1987)
16	Wiggins & Kohen (1971)
20	Werner et al. (1989)
21	Werner et al. (1983)

Note. nr = Study number according to Table 5 and 6.

D: Table 5

The six studies with differences in the LME components

nr	Study	r_a	G	R_s	R_e	C
3	LaDuca et al. (1988)	.66 $(.61)^z$	.84 $(.74)^z$	=	=	=
7	Ashton (1982)	.77 $(.75)^z$	.91 $(.86)^z$	=	=	=
17	Athanasou & Cooksey (2001)	=	.47 $(.44)^z$	.83 $(.75)^z$	=	=
18	Szucko & Kleinmuntz (1981)	=	.36 $(.32)^z$	=	=	=
28	Steinman & Doherty (1972)	.68 (.65)	.95 $(.85)^z$	=	=	=
29	MacGregor & Slovic (1986)	=	=	=	=	=

Note. nr = Study number according to Table 5 and 6. = data agreement, if the data does not agree, the Karelaia and Hogarth (2008) value is reported and supplemented by our value in parentheses.
^{z}Differences due to the not applied z-transformation in our study.

To summarize, if we compare our data (see D: Table 4 and 5), we have an agreement of 88%.

APPENDIX E: PSYCHOMETRIC META-ANALYSIS ACCORDING TO HUNTER AND SCHMIDT (2004)

Cumulating artefacts corrections in a psychometric meta-analysis

1) Cumulating artefacts

As already introduced, artefacts information was collected. In this step, each available artefact was considered separately (Hunter & Schmidt, 2005, p. 151).

First, the mean and the standard deviation of the corresponding attenuation factor was computed for each mentioned artefact (see chapter 4.5.2.3). Then, the available attenuation factors (e.g. Ave (a_j), Ave(b_j), see Equation E: 1) were combined by multiplication. An attenuation factor ($\overline{A}$ (A_j)) is the result.

$$\overline{A}\,(A_j) = \text{Ave}(a_i)*\text{Ave}(b_j)*\text{Ave}(c_j)\text{….etc.} \qquad \text{(E: 1)}$$

2) Correction of the mean correlation

In this second step, the fully corrected mean correlation ($\overline{R}$) is the corrected mean correlation in a bare-bones meta-analysis ($\bar{r}$, see Equation 2) is divided by the attenuation factor, as can be see in the following Equation E: 2:

$$\overline{R} = \text{Ave}(\rho) = \frac{\bar{r}}{\overline{A}} \qquad \text{(E: 2)}$$

3) Correcting the standard deviation of correlations

In the third step, we estimated the variance in the corrected correlation due to artefact variance. Therefore, we computed the sum of the squared coefficient of variation (*V*) across the attenuation factors (see Equation E: 3):

$$V = \frac{SD(a)^2}{Ave(a)^2} + \frac{SD(b)^2}{Ave(b)^2} + \ldots \qquad \text{(E: 3)}$$

Furthermore, we estimated the variance (*S*) in corrected study correlations, accounted for by variation in artefacts as a product (see Equation E: 4).

$$S^2 = \overline{R}^2 \overline{A}^2 V \qquad \text{(E: 4)}$$

Finally, the unexplained residual variance (S_1^2) in the corrected study correlation was calculated (see Equation E: 5):

$$S_1^2 = \overline{R}^2 - S^2 \qquad \text{(E: 5)}$$

Consequently, the fully corrected variance (Var(ρ$_j$)) is (see Equation E: 6):

$$\text{Var}(\rho_j) = \frac{S_1^{\,2} - S^2}{\overline{A}^2} \qquad \text{(E: 6)}$$

It is important to note that in the following psychometric procedures the estimation of credibility intervals, the 75% rule, and finally, the detection of moderator variables is the same as in a bare-bones meta-analysis, consequently the same steps as already reported are used.

In the following Table E: 1 represents the introduced correction of dichotomized variables according to Hunter and Schmidt (2004, see also chapter 4.5.2.3.2).

E: Table 1

The correlations corrected for dichotomizing

	Corrected correlation (Correlation according to Szucko & Kleinmuntz, 1981)				
	Components				
Judge	r_a	G	R_s	R_e	C
1	.02(.02)	-.20(-.17)	.56(.47)	.62(.52)	.11(.09)
2	.28(.23)	.20(.17)	.53(.44)	.62(.52)	.30(.25)
3	.52(.43)	.70(.58)	.59(.49)	.62(.52)	.44(.37)
4	.32(.27)	.22(.18)	.66(.55)	.62(.52)	.37(.31)
5	.40(.33)	.41(.49)	.61(.51)	.62(.52)	.36(.30)
6	.10(.08)	.91(.76)	.44(.37)	.62(.52)	-.10(-.08)
Overall	.28(.23)	.38(.32)	.56(.47)	.62(.52)	.25(.21)

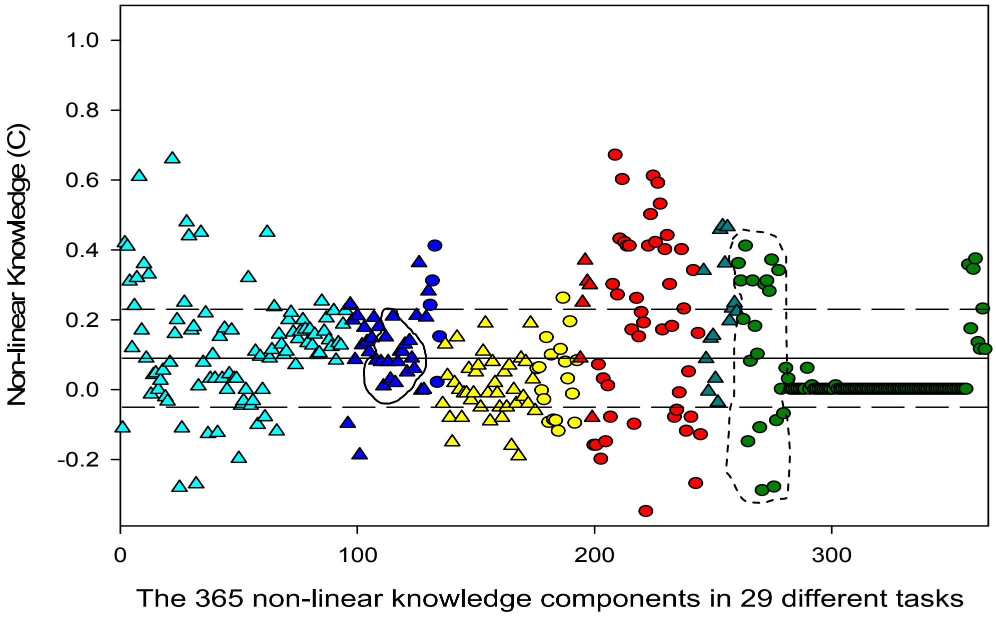
APPENDIX F: RESULTS OF OUR IDIOGRAPHIC-BASED META-ANALYSIS

F: Table 1

Judgment achievement separated into low, medium, and high level – reported by number and percent

	Judgment achievement: N (%)		
Research area	Low (>.29)	Medium (>.49)	High (<.49)
Medical science	60 (63)	13 (13)	22 (23)
Business science	17 (42)	5 (13)[a]	18 (45)
Educational science	9 (15)	9 (15)	40 (69)
Psychological science	35 (61)	16 (28)	6 (11)
Miscellaneous	59 (49)	26 (21)	35 (30)
Overall	180 (49)	69 (17)	121 (33)
Experts (210)	96 (46)	28 (13)	86 (41)
Non-experts (160)	84 (52)	41 (26)	35 (22)

Note. % = is rounded. [a]only students included



Legend

△	Medical science (experts)
▲	Business science (experts)
●	Business science (students)
△	Educational science (experts)
○	Educational science (students)
▲	Psychological science (experts)
●	Psychological science (students)
▲	Miscellaneous research areas (experts)
●	Miscellaneous research areas (students)
———	Averaged mean
— — —	80% Credibility Interval
⬭	Study with the highest number of cues (Roose & Doherty, 1976)
⬭	Study with the fewest number of cues (Steinmann & Doherty, 1972)

F: Figure 1. The scatter plot of the non-linear knowledge component (*C*) in the 365 analyzed judgments in 29 different tasks, separated into the applied research areas. The 29 different tasks are in the same order as listed in Table 5 and 6.

F: Table 2

Experts' intercorrelation of the LME components in the different areas

Components in:			Components		
Medical science	r_a	G	R_s	R_e	C
r_a	--	.85**	.14	.79**	.47**
G	.85**	--	.22*	.60**	.16
R_s	.14	.22*	--	.14	-.08
R_e	.79**	.60**	.14	--	.31**
C	.47**	.16	-.08	.31**	--
Business science					
r_a	--	.96**	.64**	.96**	.11
G	.96**	--	.49**	.95**	.11
R_s	.64**	.49**	--	.56**	-.12
R_e	.96**	.95**	.56**	--	.11
C	.11	.11	-.12	.11	
Education science					
r_a	--	.47**	.49**	.24	.24
G	.47**	--	-.16	-.44**	.00
R_s	.49**	-.16	--	.23	-.35*
R_e	.24	-.44**	.23	--	-.15
C	.24	.00	-.35*	-.15	--
Psychology science					
r_a	--	.36	.55	-.20	.87*
G	.36	--	-.41	[a]	-.14
R_s	.55	-.41	--	[a]	.81
R_e	-.20	[a]	[a]	--	[a]
C	.87*	-.14	.81	[a]	--
Miscellaneous					
r_a	--	.72**	.89**	.99**	.87**
G	.72**	--	.79**	.65**	.60*
R_s	.89**	.79**	--	.83**	.87**
R_e	.99**	.65**	.83**	--	.81**
C	.87**	.60*	.87**	.81**	--

Note. ** Correlation is significant at the .001 level (2-tailes).

* Correlation is significant at the .005 level (2-tailes).

[a] Cannot be computed because at least one of the variables is constant.

F: Table 3

Students' intercorrelation of the LME components in the different areas

Components in:	Components				
Business science	*r_a*	*G*	*R_s*	*R_e*	*C*
r_a	--	.33	-.24	[a]	.27
G	.33	--	-.56	[a]	-.82
R_e	-.24	-.56	[a]	[a]	.38
R_s	[a]	[a]	--	--	[a]
C	.27	-.82	.38	[a]	--
Education science					
r_a	--	.94**	.64**	[a]	.00
G	.94**	--	.50*	[a]	-.18
R_s	.64**	.50*	--	[a]	-.28
R_e	[a]	[a]	[a]	--	[a]
C	.00	-.18	-.28	[a]	--
Psychology science					
r_a	--	.46**	.19	.17	.26
G	.46**	--	.42**	-.67	-.30*
R_s	.19	.42**	--	-.43**	-.45
R_e	.17	-.67	-.43**	--	.44**
C	.26	-.30*	-.45**	.44**	--
Miscellaneous					
r_a	--	.93**	.64**	-.50**	.61**
G	.93**	--	.45**	-.48**	.47**
R_s	.64**	.45**	--	-.35**	.34**
R_e	-.50**	-.48**	-.35**	--	-.30**
C	.61**	.47**	.34**	-.30**	--

Note. ** Correlation is significant at the .001 level (2-tailes).

* Correlation is significant at the .005 level (2-tailes).

[a] Cannot be computed because at least one of the variables is constant.

APPENDIX G: RESULTS OF OUR NOMOTHETIC-BASED META-ANALYSIS

G: Table 1

Bare-bones meta-analysis according to the method of Hunter-Schmidt (2004) supplemented by a trim-and-fill analysis of the nonlinear knowledge component (C), separated into research area and experience level

Research area	*k*	*N*	*C*	*var$_{Corr}$*	80% CI		75%
Medicine	10	258	.19	.00	.19	.19	268.01
Business	8/10	215/221	.07/.06	.00/.00	.07/.06	.07/.06	1201.17/1285.76
Education	4	156	.02	.00	.02	.02	3999.13
Psychology	9/13	105/141	.00/-.04	.00/.00	.00/-.04	.00/-.04	959.64/769.29
Miscellaneous	12/16	249/287	.04/.00	.00/.00	.04/.00	.04/.00	361.89/260.87
Overall	43/51	983/1075	.08/.04	.00/.00	.08/.04	.08/.04	339.51/221.19
Experts in:							
Business	6	116	.08/.08	.00/.00	.08/.08	.08/.08	1216.97/1216.97
Education	2	40	.02	.00	.02	.02	124434
Psychology	4/6	59/70	-.04/-.06	.00/.00	-.04/-.06	-.04/-.06	628.52/601.52
Miscellaneous	5/6	15/23	.22/.08	.00/.00	.22/.08	.22/.08	2872.94/869.40
Overall[a]	27/28	488/554	.12/.07	.00/.00	.12/.07	.12/.07	378.19/219.50
Students in:							
Business	2	99	.05	.00	.05	.05	1677.99
Education	2	116	.02	.00	.02	.02	1677.89
Psychology	5/7	46/62	.04/.06	.00/.00	.04/.06	.04/.06	3314.43/4019.04
Miscellaneous	11	234	.03/-.03	.00/.00	.03/-.03	.03/-.03	506.97/248.42
Overall	20	495	.03/.00	.00/.00	.03/.00	.03/.00	710.93/322.24

Note. *k* = number of correlations (i.e. judgment tasks). *N* = total sample size for all judgment tasks combined. *C* = weighted mean correlation according to Hunter and Schmidt (2004). var$_{corr}$ = corrected variation according to Hunter and Schmidt (2004, variance of true score correlation). 80% CI = 80% credibility interval for true score correlation distribution. 75% = Percentage variance of observed correlations due to all artefacts, if below 75%, it indicates moderator variable. [a]this analysis includes medical experts. /Results of the trim-and-fill analyses after a publication bias is indicated.

G: Table 2

Psychometric meta-analysis of the component (C) in different research areas, separated by experience levels

				Overall			*Experts*			*Students*		
Research area	*rr*	*k (experts)*	*N (experts)*	*C*	var_{corr}	75%	*C*	var_{corr}	75%	*C*	var_{corr}	75%
Medical science		10	258	**.25**	**.00**	**271.75**	**.25**	**.00**	**271.75**	[a]	[a]	[a]
Business science		8(6)	215(116)	**.08** (.07)	**.00** (.00)	**1201.17** (1285.76)	**.10** (.09)	**.00** (.00)	**1216.97** (1329.88)	**.06**	**.00**	**1677.99**
Education science	.90	4(2)	156(40)	**.03**	**.00**	**3348.19**	**.03**	**.00**	**> 10000**	**.04**	**.00**	**1692.48**
	.78			**.04**	**.00**	**3347.46**	**.03**	**.00**	**> 10000**	**.05**	**.00**	**1691.03**
	.50			**.06**	**.00**	**3346.11**	**.05**	**.00**	**> 10000**	**.07**	**.00**	**1686.72**
Psychology	.90	9(4)	105(59)	**.00** (-.05)	**.00** (.00)	**959.64** (769.29)	**-.04** (-.06)	**.00** (.00)	**628.53** (601.51)	**.05** (.06)	**.00** (.00)	**3314.43** (4019.04)
	.78			**.00** (-.05)	**.00** (.00)	**959.64** (769.29)	**-.04** (-.07)	**.00** (.00)	**628.53** (601.51)	**.05** (.07)	**.00** (.00)	**3314.43** (4019.04)
	.50			**-.01** (-.08)	**.00** (.00)	**959.64** (769.29)	**-.05** (-.09)	**.00** (.00)	**628.53** (601.51)	**.09** (.11)	**.00** (.00)	3314.43 (4019.04)
Miscellaneous	.90	12(5)	249(15)	**.04** (.00)	**.00** (.00)	**361.89** (260.87)	**.23** (.08)	**.00** (.00)	**2872.94** 869.41	**.03** (-.03)	**.00** (.00)	**506.97** 248.42
	.78			**.05** (.00)	**.00** (.00)	**361.89** (260.87)	[b]	[b]	[b]	**.04** (-.04)	**.00** (.00)	**506.97** 248.42
	.50			**.08** (.00)	**.00** (.00)	**361.89** (260.87)	[b]	[b]	[b]	**.06** (-.07)	**.00** (.00)	**506.97** 248.42
Overall	.90	43(27)	983(488)	**.10** (.05)	**.00** (.00)	**340.42** 216.39	**.15** (.15)	**.00** (.00)	**379.50** 361.70	**.03** (.00)	**.00** (.00)	**710.93** (322.24)
	.78			**.10** (.06)	**.00** (.00)	**340.36** 216.37	**.16** (.15)	**.00** (.00)	**379.38** 361.59	**.04** (.00)	**.00** (.00)	**710.93** (322.24)
	.50			**.13** (.06)	**.00** (.00)	**341.15** 216.52	**.17** (.17)	**.00** (.00)	**380.55** 362.66	**.07** (.00)	**.00** (.00)	**711.04** (322.24)

Note. Values enclosed in parentheses represent our results of the trim-and-fill method application if a publication bias is indicated**.** *rr* = retest-reliability values used in our measurement-error corrections. *k* = Number of correlations according to Hunter and Schmidt (2004). *N* = Total sample size according to Hunter and Schmidt (2004). *C* = mean true score correlation according to Hunter and Schmidt (2004). var_{corr} = corrected variation according to Hunter and Schmidt (2004, variance of true score correlation). 75% = Percentage variance of observed correlations due to all artefact, if below 75%, it indicates moderator variable. [a]only experts are included in the medical science category. [b]no further correction because only meteorologists are included.

G: Table 3

Experts' intercorrelation of the LME components in the different areas

Components in:			Components		
Business science	*r_a*	*G*	*R_s*	*R_e*	*C*
r_a	--	.95**	.27	.96**	.39
G	.95**	--	.24	.90*	.65
R_s	.27	.24	--	.34	-.25
R_e	.96**	.90*	.34	--	.34
C	.39	.65	-.25	.34	--
Education science					
r_a	--	-1.00**	1.00**	1.00**	1.00**
G	-1.00**	--	-1.00**	-1.00**	-1.00**
R_s	1.00**	-1.00**	--	1.00**	1.00**
R_e	1.00**	-1.00**	1.00**	--	1.00**
C	1.00**	-1.00**	1.00**	1.00**	--
Psychology science					
r_a	--	.99*	-.91	-.78	.68
G	.99*	--	-.88	-.72	.56
R_s	-.91	-.88	--	.96*	-.83
R_e	-.78	-.72	.96*	--	-.93
C	.68	.56	-.83	-.93	--
Miscellaneous					
r_a	--	.89*	.88*	.99**	.94*
G	.89*	--	.99	.82	.94*
R_s	.88**	.99**	--	.81	.95*
R_e	.99**	.82	.81	--	.90*
C	.94*	.94*	.95*	.90*	--

Note. ** Correlation is significant at the .001 level (2-tailes).

* Correlation is significant at the .005 level (2-tailes).

G: Table 4

Students' intercorrelation of the LME components in the different areas

Components in:			Components		
Business science	r_a	G	R_s	R_e	C
r_a	--	.97	.92	1.00*	1.00**
G	.97	--	.99	.94	-1.00**
R_s	.92	.99	--	.89	-.100**
R_e	1.00*	.94	.89	--	1.00**
C	1.00**	-1.00**	-1.00**	1.00**	--
Education science					
r_a	--	1.00**	-1.00**	-1.00**	-1.00**
G	1.00**	--	-1.00**	-1.00**	-1.00**
R_s	-1.00**	-1.00**	--	1.00**	1.00**
R_e	-1.00**	-1.00**	1.00**	--	1.00**
C	-1.00**	-1.00**	1.00**	1.00**	--
Psychology science					
r_a	--	-.07	1.00**	.14	-.13
G	-.07	--	-.07	.86	-.94*
R_s	1.00**	-.07	--	.14	-.13
R_e	.14	.86	.14	--	-.85
C	-.13	-.94*	-.13	-.85	--
Miscellaneous					
r_a	--	.81**	.94**	.22	.26
G	.81**	--	.72*	-.24	-.24
R_s	.94**	.72*	--	.03	.38
R_e	.21	-.24	.03	--	.53
C	.26	-.24	.38	.53	--

Note. ** Correlation is significant at the .001 level (2-tailes).

* Correlation is significant at the .005 level (2-tailes).

[a] Cannot be computed because at least one of the variables is constant.

APPENDIX H: RESULTS OF OUR ROBUSTNESS ANALYSIS

H: Table 1

Judgment achievement (r_a) estimated by the fixed-effect model and by a random-effect model

Model	r_a	*SE*	95% CI
Research area			
Medicine			
FE	.39	.06	.27 - .51
RM	.39	.06	.27 - .51
Business			
FE	.49	.06	.37 - .62
RM	.50	.12	.26 - .74
Education			
FE	.38	.08	.23 - .54
RM	.38	.08	.23 - .54
Psychology			
FE	.22	.06	.09 - .34
RM	.22	.06	.09 - .34
Miscellaneous			
FE	.44	.06	.31 - .56
RM	.47	.07	.33 - .62
Overall			
FE	.38	.03	.33 - .44
RM	.39	.03	.32 - .46

Note. r_a = weighted mean correlation. 95% CI = confidence interval. FE = Fixed-effect model. RM = Random-effect model (DerSimonian & Laird, 1986)

H: Table 2

Knowledge component (G) estimated by the fixed-effect model and by a random-effect model

Model	*G*	*SE*	95% CI
Research area			
Medicine			
FE	.60	.06	.48 - .72
RM	.60	.06	.46 - .73
Business			
FE	.66	.06	.53 - .79
RM	.66	.11	.43 - .87
Education			
FE	.73	.08	.57 - .88
RM	.73	.08	.57 - .88
Psychology			
FE	.38	.09	.18 - .56
RM	.41	.11	.18 - .63
Miscellaneous			
FE	.68	.06	.55 - .80
RM	.77	.09	.58 - .96
Overall			
FE	.63	.03	.57 - .69
RM	.64	.04	.55 - .73

Note. G = weighted mean correlation. 95% CI = confidence interval. FE = Fixed-effect model. RM = Random-effect model (DerSimonian & Laird, 1986)

H: Table 3

Consistency component (R_s) estimated by the fixed-effect model and by a random-effect model

Model	R_s	*SE*	95% CI
Research area			
Medicine			
FE	.80	.06	.68 - .93
RM	.80	.06	.68 - .93
Business			
FE	.80	.06	.67 - .93
RM	.80	.06	.67 - .93
Education			
FE	.73	.08	.57 - .88
RM	.73	.08	.57 - .88
Psychology			
FE	.78	.08	.62 - .94
RM	.78	.08	.62 - .94
Miscellaneous			
FE	.71	.06	.58 - .83
RM	.71	.06	.58 - .83
Overall			
FE	.76	.03	.71 - .82
RM	.76	.03	.71 - .82

Note. R_s = weighted mean correlation. 95% CI = confidence interval. FE = Fixed-effect model. RM = Random-effect model (DerSimonian & Laird, 1986)

H: Table 4

Environmental predictability (R_e) estimated by the fixed-effect model and by a random-effect model

Model	R_e	*SE*	95% CI
Research area			
Medicine			
FE	.66	.06	.54 - .79
RM	.66	.06	.54 - .79
Business			
FE	.70	.06	.58 - .83
RM	.71	.06	.58 - .83
Education			
FE	.70	.08	.54 - .86
RM	.70	.08	.54 - .86
Psychology			
FE	.68	.06	.56 - .80
RM	.68	.06	.56 - .80
Miscellaneous			
FE	.88	.06	.76 - 1.00
RM	.88	.06	.75 - 1.00
Overall			
FE	.73	.03	.67 - .78
RM	.73	.03	.67 - .78

Note. R_e = weighted mean correlation. 95% CI = confidence interval. FE = Fixed-effect model. RM = Random-effect model (DerSimonian & Laird, 1986)

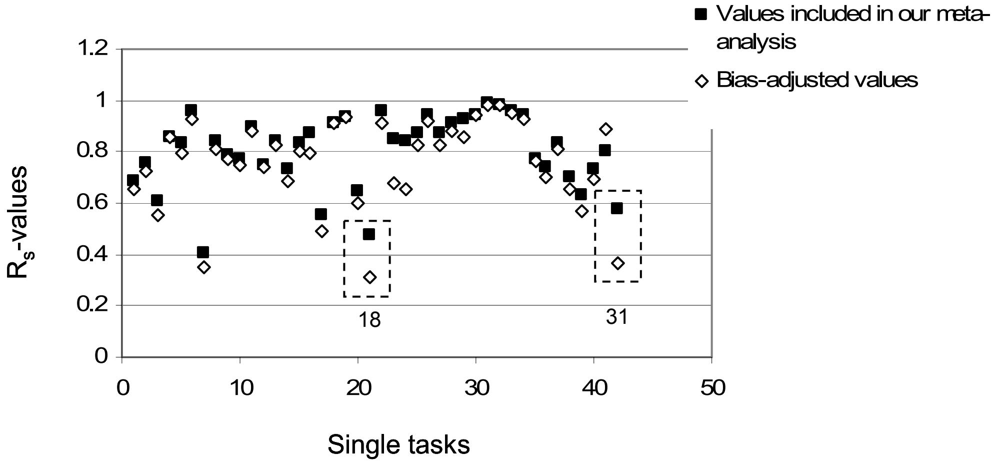
H: Table 5

Non-linear knowledge component (C) estimated by the fixed-effect model and by a random-effect model

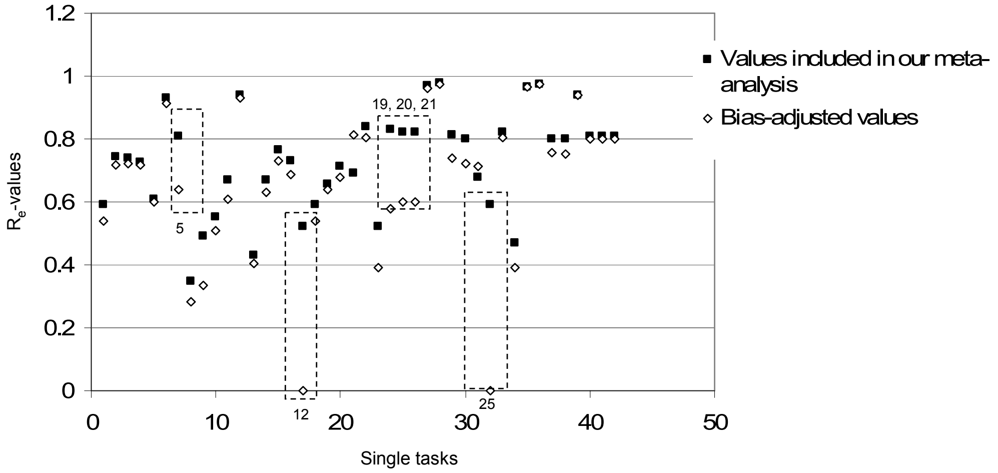
Model	*C*	*SE*	95% CI
Research area			
Medicine			
FE	.18	.06	.06 - .30
RM	.18	.06	.06 - .30
Business			
FE	.07	.06	-.06 - .20
RM	.07	.06	-.06 - .20
Education			
FE	.02	.08	-.13 - .18
RM	.02	.08	-.13 - .18
Psychology			
FE	-.00	.09	-.19 - .18
RM	-.00	.09	-.19 - .18
Miscellaneous			
FE	.05	.07	-.09 - .20
RM	.05	.07	-.09 - .20
Overall			
FE	.08	.03	.02 - .15
RM	.08	.03	.02 - .15

Note. C = weighted mean correlation. 95% CI = confidence interval. FE = Fixed-effect model. RM = Random-effect model (DerSimonian & Laird, 1986)

APPENDIX I: BIAS-ADJUSTED R^2



I: Figure 1. Comparison of R_s bias-adjusted values and non-adjusted values included in our meta-analysis.



I: Figure 2. Comparison of R_e bias-adjusted values and non-adjusted values included in our meta-analysis.

Legend

⬚ Studies with great differences between values included in our meta-analysis and bias-adjusted values. These studies are labeled by their study number see Tables 5, 6.

I: Table 1

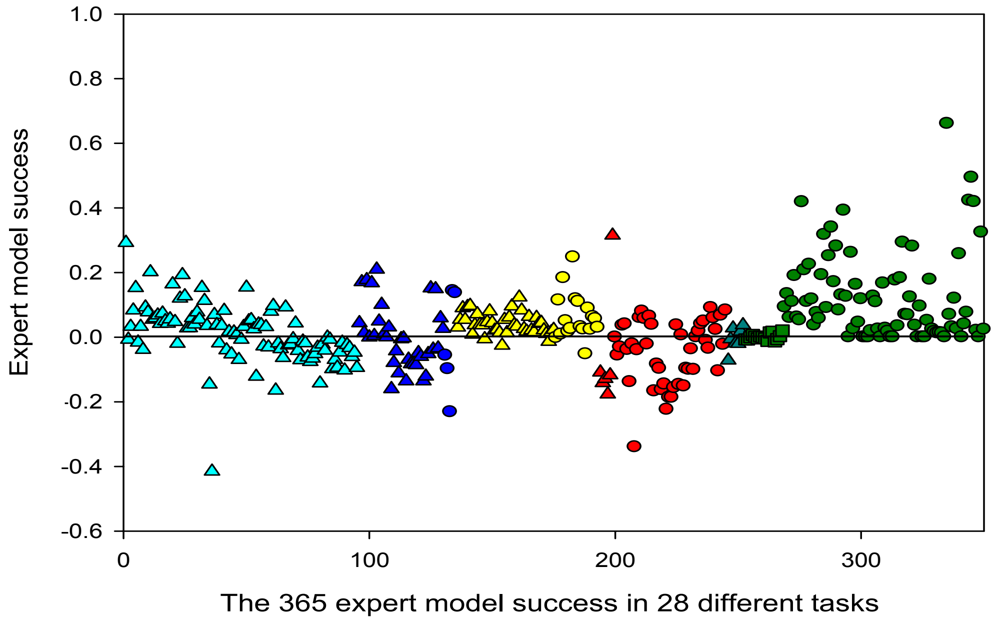
Meta-analysis according to Hunter and Schmidt (2004).

Meta-analysis	*k*	*N*	*r*	SD_{ra}	*95% CI*		*Q*
Non-corrected							
R_s- values	39[1]	1007	.77	.01	.73	.80	79.69***
Bias-adjusted R_s- values	39[1]	1007	.72	.01	.67	.77	98.20***
Non-corrected							
R_e- values	41[1]	979	.72	.02	.67	.77	106.27***
Bias-adjusted R_e- values	41[1]	979	.67	.03	.61	.73	126.01***

Note. k = number of correlations (i.e. judgment tasks); *N* = total sample size for all judgment tasks combined; $\underline{r_a}$ = average corrected correlation according to Hunter and Schmidt (2004); SD_{ra} = Standard deviation of corrected correlation according to Hunter and Schmidt (2004); SD_{res} = residual standard deviation; *95% CI* = 95% confidence interval; *Q* = statistic used to test for homogeneity in the true correlations across judgment tasks; *** $p < .001$.[1] three judgment tasks were excluded (Einhorn, 1974; Kim et al., 1987) because it was not possible with the Wright syntax (2005) to include tasks with only three judges.

Although there are some differences indicated, our analysis shows that if the bias-adjusted correction would influence our results then psychological values are rather overestimated then underestimated.

APPENDIX J: SUCCESS OF SINGLE EXPERT MODELS



J: Figure 1. The scatter plot of single expert model success (GR_e-r_a).

Note. The legend you will find on page XV.

According to Camerer (1981) and Goldberg (1970) with the product of the lens model components knowledge (*G*) and environmental predictability (R_e) the validity of the expert model (i.e. regression model, *LME*) are captured. As research has shown, often judgments based on the perfectly reliable regression model perform better then the original judgment by the less than perfectly reliable human. Therefore, it can also be shown how well the regression model, or simply a linear model, substitutes the judge as measure of expert success by subtracting judgment achievement from the product term (GR_e, see Camerer, 1981, p. 413).

However, as our scatter plots imply high heterogeneity this should be the scope of further research to reveal some regularity. For example, can the expert model success in educational and other research areas be confirmed with the nomothetic data base?