LUNDS UNIVERSITET
Institutionen
för psykologi

Does correction for guessing improve the realism in absolute and relative frequency judgements?

Tom-Olof Romu

Psykologexamensuppsats VT2005

Supervisors:
Carl Martin Allwood
Marcus Johansson

Examinator:
Sven Ingmar Andersson

Romu, T-O. *Does correction for guessing improve the realism in absolute and relative frequency judgements?* Psykologexamensuppsats. Institutionen för psykologi, Lunds universitet. Vol. VII (2005): 08

Abstract

The present study investigated the effects of three factors on the level of realism in frequency judgements. These factors were: Instruction (No correction for guessing/Correction for guessing), Format (whether the frequency judgements were made in terms of absolute numbers or in percent) and Difficulty level (easy/hard set of questions). All three factors were found to have a significant effect on the level of the frequency judgements and their realism. In addition, there were no interaction effects. The results suggest that the level and the realism of frequency judgements are both affected by different factors. Correction for guessing improved the realism in frequency judgements, however, only on the set of easy questions and markedly only in the relative format.

Key words: calibration research, confidence and frequency judgements, metacognition, realism

Acknowledgements

I would like to thank, from the depth of my heart, my supervisor *Carl Martin Allwood* for his gentle encouragement, insightful critique, constructive suggestions and engagement in the process of writing this paper. If it wasn't for him, I would never have been able to make heads or tails out of the results of the analysis of the data. Another source of inspiration was *Marcus Johansson*, whose creative mind in fact was responsible for outlining the design of this study and should be credited accordingly. He was further very helpful in the preliminary adaptation of the data. Many sincere thanks go to him. Finally, I would like to thank all the lecturers and participants who gave up a bit of their precious time to help me in this investigation.

Contents

Does correction for guessing improve the realism in absolute and relative frequency judgements?

Confidence and frequency judgements are part of what cognitive psychologists refer to as metacognition. In short, a confidence judgement is a statement about how sure one is that the answer one has selected to a question is correct and a frequency judgement is a statement about how many questions in a test one believes to have answered correctly. Metacognition is usually described as encompassing both knowledge about one's cognitive abilities and regulation of processes that coordinate cognition, in short, cognition about cognition (Fernandez-Duque, Baird, & Posner, 2000; Flavell, Miller, & Miller, 1993; Koriat, 2000; Shimamura, 2000).

To illustrate, Nelson (1996) described some metacognitive components in connection to learning and distinguished between control and monitoring components. During the acquisition phase the key metacognitive control components are the selection of a strategy as to which mnemonic processing is best suited for the task and the allocation of study time. The key metacognitive monitoring components are an assessment of previous knowledge, a judgement of learning concerning how well the material has been memorized and judgements of confidence for material retrieved from memory, that is, how sure one is that the retrieved answer is correct.

In this context it is interesting to note the parallel to the emerging cognitive neuroscience of metacognition (Banich, 1997). Shimamura (2000) has suggested that "…there is considerable convergence of issues associated with metacognition, executive control, working memory and frontal lobe function" (p. 313).

There are several other types of judgements that can be labelled metacognitive, for instance information-based judgements of the extent of one's knowledge (Allwood & Granhag, 1996a; Rozenblit & Keil, 2002), experience-based judgements (i.e. an immediate "gut feeling") (Koriat, 2000; Koriat & Levy-Sadot, 2000) and judgements based on the distinction between recollection and familiarity which has some relation to the tip-of-the-tongue phenomenon, which can be seen as an example of the feeling of knowing (Dodson & Schacter, 2002; Koriat, 2000, Yonelinas, 2001, 2002).

One aspect of metacognitive judgements is that they can be more or less valid, that is, *realistic* compared to some objective criterion (e.g. an agreed upon answer). There are several quite different accounts of which factors affect the realism in people's confidence judgements (for overviews, see Allwood & Granhag, 1999; Keren, 1991; Lichtenstein, Fischoff, & Phillips, 1982; McClelland & Bolger, 1994; Suantak, Bolger, & Ferrell, 1996).

One reasonable assumption is that people's metacognition is influenced by a multitude of factors, for instance anchoring (Mussweiler & Strack, 2000; Scheck, Meeter, & Nelson, 2004;

Tversky & Kahneman, 1974), bias (Mussweiler & Neumann, 2000); problem representation (Cosmides & Tooby, 1996), problematizing (Allwood, 1995) and self-evaluation (Dunning, Johnson, Ehrlinger & Kruger, 2003; Rozenblit & Keil, 2002). In the present study I investigate the influence of some different factors on the level of realism in frequency judgements.

In the last decades there have been forthcoming several empirically based accounts of different processes that have been shown to affect metacognitive judgements. Some of these are social factors (Johansson, 2004), rule-based versus exemplar-based processing (Juslin & Olsson, 2004), heuristics (Kahneman & Tversky, 1996), ecological rationality (Todd & Gigerenzer, 2003) and a multitude of other factors as, for instance, type of feedback, type of confidence judgement and the temporal relation between task, decision and outcome (Allwood & Granhag, 1999). Concerning the specifics of frequency judgements, (also called aggregated-item judgements, Treadwell & Nelson, 1996; and global judgements, Liberman, 2004; Sniezek & Buckley, 1991) there has been much less research.

The main finding within this latter research is the *confidence-frequency effect* which means that when confidence judgements of answers to general knowledge questions (hereafter GKQs) result in overconfidence then frequency judgements typically result in good realism, and when confidence judgements show good realism then frequency judgements typically show underconfidence. There are only few explanations why this effect occurs. For example, Sniezek and Buckley (1991) proposed a dual-process account which states that confidence and frequency judgements are made with reference to different considerations. They propose that item-specific confidence judgements may be affected by memory (i.e. memory-cues and memory distortions) and memory processes such as the distinction between recollection and familiarity and item-specific considerations whereas frequency judgements likely are influenced by considerations of one's expertise, previous task performances and so forth.

Liberman (2004) has suggested that the confidence-frequency effect can be diminished and manipulated depending on what instructions the participants in different experimental conditions receive. He presented empirical support for his hypothesis that when making a frequency judgement participants do not consider that on all items they claimed they were guessing, they may be expected to be correct on half of them just by chance and should incorporate this in their frequency judgement.

Research questions within the field of metacognitive judgements

Before we turn our attention to the specifics of investigating what factors affects the realism in frequency judgements I will present a general background. Several lines of research within

psychology such as judgement and decision making, learning and memory research come together under the concept of metacognition and influence the tradition of what is called *calibration research*. The specific focus within calibration research is on how well people's subjective confidence and frequency judgements conform to an objective measure of their memory or knowledge statements.

Studies of calibration have shown that people's confidence often exceeds their accuracy. Overconfidence is common, albeit not universal, since it can be reversed for easy questions. (For reviews of the literature, see Keren, 1991; Lichtenstein et al., 1982, McClelland & Bolger, 1994). This phenomenon, called the difficulty effect or the *hard-easy effect*, is the typical finding that increasing overconfidence often co-varies with increasing task difficulty or lower proportions of correct answers to GKQs. Incidentally, in contrast to this, underconfidence (i.e. being less confident than correct) has been observed for perceptual tasks such as sensory discrimination (Juslin & Olsson, 1997; Juslin, Winman, & Olsson, 2003; Olsson & Juslin, 2000). The main challenge facing theories and models of calibration is to explain these effects.

Another facet of metacognition concerns memory research. Numerous multiple-memory systems have been proposed over the years such as, to name a few, declarative memory versus procedural memory, explicit versus implicit memory, episodic versus semantic memory and working memory versus reference memory (Banich, 1997; Galotti, 1999; Kellogg, 1995).

The distinction sometimes made in memory research between remembering and knowing is assumed to tap into the different parts of the episodic and semantic memory systems. This distinction of remembering versus knowing also designates the difference between recollection and familiarity which are argued to be two separate processes that underlie recognition memory and there is sometimes an opposition relationship between familiarity and item-specific memory as when people see a familiar test item but fail to remember more specific information about it (Dodson & Schacter, 2002; Yonelinas, Dobbins, Szymanski, Dhaliwal, & King, 1996). One could reasonably argue that when people are engaged in tasks that require them to make confidence and frequency judgements they make reference to different parts of their memory systems as in Sniezek and Buckley's (1991) dual-process account mentioned earlier. In this work I used declarative general knowledge tasks which could be affected by both of the above mentioned processes.

In general, research on metacognition by cognitive scientists and neurobiologists can help produce a "…synergy between the psychological and philosophical approaches to consciousness, by furnishing constraints on the range of acceptable theories and by providing clues to inspire new theories" (Nelson, 1996, p. 103). We have come a long way since the 19th century

philosopher Comte formulated his famous paradox; "The thinker cannot divide himself into two, of whom one reasons whilst the other observes him reason. The organ observed and the organ observing being, in this case, identical, how could such observation take place?" (quoted in Nelson, 1996, p. 104). One way to dissolve this paradox of self-reference was developed by the mathematician and logician Alfred Tarski who came up with the idea of the meta concept which is a level up from the object level and is in some sense separable from the object level it refers to (Popper, 1963, 1990). His solution was that no sentence could refer to itself. His theory of truth made use of semantics to inform logic but the same reasoning applied to consciousness informs us that at the object-level there are cognitions about for instance external objects and at the meta-level there are cognitions concerning cognitions of those external objects. They are two possibly different processes that could occur simultaneously but may occur on different levels. However, the exact meaning of "level" has often been left unspecified. In spite of this, today, this distinction has been a useful heuristic and has been further differentiated in for instance research on the neural bases of consciousness and metacognition (Freeman, 1999) and it has also been used to introduce new theories about the workings of the mind (Koriat & Levy-Sadot, 2000; Schooler, 2002).

The different lines of research concerning metacognition briefly delineated above illustrate that there are many factors intertwined in calibration research. How to tease them apart or understand how they interact when forming a metacognitive judgement is no easy task for the researcher.

The aim and scope of this paper

The aim of this work is to achieve an improved understanding of some of the factors that affect frequency judgements. As noted above, previous research has often found that frequency judgements are given at a lower level than item specific confidence judgements. One recent explanation of this effect is attempted by Liberman (2004) who presented empirical evidence for that the level of frequency judgements can approach the level of confidence judgements if the participants were given instructions to incorporate a correction for random guessing. I wanted to see if this effect can be replicated in the present study.

Furthermore, it has been suggested in a previous study by Brenner, Koehler, Liberman, and Tversky (1996) that relative frequency judgements (i.e., answers given as percentage correct of a set), as compared with absolute frequency judgements (i.e., answers given in natural numbers) might exhibit overestimation. The researchers mentioned did not actually perform an experiment to investigate if this actually is true. I wanted to find out if that suggestion holds with respect to

frequency judgements. Finally, I investigate if the level of difficulty of the questions answered has a significant effect on the level of frequency judgements and on their realism. In brief, this study investigates if certain conditions may have a conducive effect on frequency judgements so they approach the actual accuracy of answers given.

Theories and models of calibration

As mentioned above, the main challenge facing theories and models of calibration is to provide an explanation to the overconfidence effect, the hard-easy effect (Juslin et al., 2000; Keren, 1991; Klayman, Soll, González-Vallejo, & Barlas, 1999; Lichtenstein et al., 1982; McClelland & Bolger, 1994; Suantak et al., 1996) and more recently, the confidence-frequency effect (Brenner et al., 1996; Griffin & Buehler, 1999; Treadwell & Nelson, 1996). Over the years numerous theoretical accounts and models have been developed to explain these effects. This section is intended as a selective overview to provide the reader with the core assumptions of some of the major models of calibration and how their explanations differ concerning the overconfidence and the hard-easy effect. Note that only some of the models also provide an explanation of the confidence-frequency effect which will be dealt with later when I present previous research on frequency judgements.

At least since the 90'ies, there have been two major camps in calibration research, each with "a radically different view" (McClelland & Bolger, 1994, p. 455) as to where the locus of the observed biases in calibration should be located. One camp has largely put it within the individual. The most representative members of this internalist or subjectivist camp are Daniel Kahneman and Amos Tversky. In propounding their "heuristics and biases program" (Tversky & Kahneman, 1974) they have provided numerous demonstrations of the apparent irrationality of individuals when they engage in probabilistic reasoning and making judgements under uncertainty. Kahneman and Tversky have said that their research results are discouraging for those who "… wish to view man as a reasonable intuitive statistician" (cited in Goldman, 1986, p. 305). Several other researchers (e.g. Keren, 1991) have similarly built their explanations on one or more internal heuristics to explain the relationship between difficulty and over/underconfidence.

Keren (1991) suggested that participants use the "anchor and adjust" heuristic on a probability estimate reflecting intermediate difficulty (75 %) on GKQs in a two-alternative forced choice (2AFC) task. According to this account, participants do not expect a task that is so difficult that performance will be on a chance level nor do they expect a task that is so easy that performance will be near perfect. If an item is perceived to be very easy or very hard one would

adjust accordingly from the preconceived anchor of intermediate difficulty, but not sufficiently and this would explain under- or overconfidence. Exactly how participants would make these adjustments is not outlined by the author.

In contrast to the internalists, the other camp might be called the externalists in that they have argued that the locus of bias in the main is found outside the cognitive system of the individual. The person who most forcefully has put forward arguments against the internalist view is, according to McClelland and Bolger (1994), Gerd Gigerenzer. He has made the claim that "biases in probabilistic reasoning are essentially artifacts" (McClelland & Bolger, 1994, p. 456), built into the tests by using a non-representative sampling of questions which is termed "misleading tasks". It should be noted that this explanation is in itself troublesome. First of all it's not obvious what items should be included or excluded to constitute a well-specified reference-class. Secondly, the argument against the use of non-representative items assumes that people have similar general background knowledge since the use of representative items in itself is argued to be enough to make the hard-easy effect disappear.

Gigerenzer (McClelland & Bolger, 1994) has also argued that people's probabilistic representations are in a frequentist format and not formed according to Bayes' rule. This will be explicated further below. For now it suffices to say that Cosmides and Tooby (1996) made a series of studies that showed, somewhat surprisingly, that people perhaps are good intuitive statisticians conforming to Bayes' rule after all, as long as the nature of the problem is presented in a frequentist format! This is an example of, as research has progressed, the above distinction between the internalist and the externalist has been blurred, which can be noticed in some models and accounts of probability judgements in that they are able to accommodate a wider range of phenomena as in for instance Support theory (Tversky & Koehler, 1994).

In short, Suantak et al. (1996) has pinpointed four factors that have been used to explain the hard-easy effect which is just mentioned briefly as a mnemonic device for the reader since they will be further delineated in connection to the various models presented in the next sections.

The hard-easy effect (1) is due to bias as in lack of attention to the quality or weight of evidence or disconfirming evidence rather than sensitivity to task difficulty as in the strength and weight model, (2) is due to response criteria in which respondent's lack of information about the difficulty of the task prevents them to adjust to varying degrees of difficulty as Keren suggested (3) is a stimulus selection artefact in which the usually reported overconfidence should, according to the ecological models, disappear when questions are representative of a given reference class and(4) is mainly due to statistical error. However, Klayman et al. (1999) have shown how one may separate systematic psychological effects from statistical effects and still find systematic

differences between confidence judgements and accuracy such as the often found hard-easy effect and the typical overconfidence phenomenon.

Internalist models

The stage model

In 1980 Koriat, Lichtenstein and Fischoff (described in McClelland & Bolger, 1994) presented a three-stage model of the cognitive processes involved in answering 2AFC GKQs to make a confidence judgement. First, memory is searched for relevant information and an answer is chosen, secondly one assesses the evidence to arrive at a feeling of certainty and thirdly this feeling is transposed to a numerical estimate. Koriat et al. suggested that overconfidence might result because of less than perfect processing in any of the three stages.

In the first stage a person might be biased in how they elicit knowledge, perhaps by favouring positive evidence. In the second stage individuals might have a tendency to disregard evidence inconsistent with the chosen answer and this would lead to overconfidence. In the last stage there could possibly occur a mistranslation of the feeling of certainty when transforming it to a numerical value. The authors themselves found some, empirical support for their model whereas subsequent research has not been able to replicate these findings (e.g. Allwood & Granhag, 1996b). Despite the lack of empirical support for their model it is useful when viewed as a framework where other models of calibration can, at least partly, be positioned as has been suggested by McClelland and Bolger (1994). For example, the process model (described next) can be located at the first stage, the strength and weight model (next after that) explains miscalibration as occurring in the first and second stages and so forth.

The process model

As outlined by McClelland and Bolger (1994) this model was developed by R. S. May in 1986 and "contain many interesting and novel ideas, some of which were taken up in the later ecological models" (p. 468). She suggested that miscalibration was a consequence of the specific background knowledge possessed by different individuals, of the type of task performed and the selection of items within the tasks. Concerning GKQs May suggested two possible ways, or types of mental models, in which a person's knowledge could be internally represented.

The first mental model is in the form of syllogisms where the individual uses inference to choose between the two alternatives in a 2AFC item. If presented with the following question "Which country has more inhabitants? (a) Sweden, (b) Netherlands." May proposed that an individual might reason along the following lines; usually the larger the country the more

inhabitants in it, Sweden has a much larger area and is therefore more likely to have more inhabitants. On this basis, the individual would choose Sweden as the correct answer. Note that the individuals' confidence judgement would be a function of her/his knowledge of nations and their respective sizes and populations and other possible background knowledge. May argued that if the items were randomly sampled from representative selections of nations then good calibration would be expected, but if there are many items which are "misleading", that is, items for which the inference would produce the wrong answer (as in the example above), then overconfidence would occur.

The second mental model is in the form of a cognitive map. This type of knowledge representation could be used to answer a question like "Which city is further north? (a) Rome, (b) New York." May showed that confidence was highly correlated with the distortions of individuals' cognitive maps. In other words, the subjective distance between the cities would determine the answer and the respective confidence given, not the objective geographical relationship.

This second mental model has received substantial support in subsequent research in neuropsychology and has been referred to as a part of "The 'where' dorsal visual system for spatial processing" (Banich, 1997, p. 204). However, May was clearly wrong, as has been shown by subsequent research (e.g. Brenner et al., 1996), that perfect calibration always and only results in the absence of "misleading" items.

The strength and weight model

As Johansson (2004) amply put it in his description of Griffin and Tversky's 1992 article on the realism in confidence judgements, "…they resort to two concepts, *strength* and *weight"* (p. 10). Even though these concepts lack rigid definitions, by "…strength they mean the 'extremeness' of available evidence and by weight the 'predictive validity' of the evidence" (McClelland & Bolger, 1994, p. 463). In short, Griffin and Tversky argue that overconfidence results when individuals excessively rely on the strength of evidence with simultaneous insufficient adjustment to its weight. As an example, individuals may use the representativeness heuristic when judging if they should trust what, for instance, a lawyer is saying by attending to how confident he or she seems whilst ignoring other factors, such as degree of experience this lawyer has in a hypothesized case. In a similar vein, underconfidence results when strength is lower than weight. They also hypothesize that individuals make use of the anchor-and-adjust heuristic but usually when adjusting to the weight of evidence they fail to adjust sufficiently. Tversky further expounded on these ideas together with Koehler with whom he developed Support theory, described next.

Support theory

Developed by Tversky and Koehler (1994), this is probably the broadest model since it tries to accommodate a wider range of phenomena than those specifically studied in calibration research, such as ordinal judgements and assessment of upper and lower probabilities. The basic features of the model have been extended further in "local-weight models for decomposition of evidential support" by Brenner and Koehler (1999) and Brenner's (2003) "random support model" which focuses on the realism in confidence judgements. To avoid the fine-grained technicalities of the latter models I restrict myself to the basic premises of the general support theory.

In order to get the "feel" of support theory I will relate an example that Tversky and Koehler use to illustrate what they suggest is representative of an essential feature of human judgement. In a study by Fischoff, Slovic, and Lichtenstein (mentioned in Tversky & Koehler, 1994) they asked car mechanics to figure out the probabilities of different causes why a car wouldn't start. They had a main hypothesis, called the focal hypothesis, but the interesting part is that they found that the mean probability assigned to the residual hypothesis, "The cause of failure is something other than the battery, the fuel system, or the engine", increased from .22 to .44 when the residual hypothesis was unpacked into more specific causes, e.g., the ignition system. Even though the participants were experienced car mechanics and could reasonably be assumed to be aware of these possibilities, the point is that they discounted the hypotheses not explicitly mentioned.

This was used by Tversky and Koehler (1994) to form a basic premise of support theory, namely, "…probability judgements are attached not to events but to descriptions of events" (p. 549). The example above illustrates that the very same hypothesis could be associated with varying degrees of confidence, depending on its description or simply because it's made salient, which in turn influences the balance of support given to the focal hypothesis and its alternative. Similarly to Griffin and Tversky's strength and weight model described above, support is defined as the strength of the evidence which could be based on known objective data on actual frequencies of the occurrence of some event or on the subjective memories of the person trying to evaluate the alternative hypotheses. These recollections are further "…mediated by judgemental heuristics, such as representativeness, availability or anchoring and adjustment" (Tversky & Koehler, 1994, p. 549). Support could also be mediated by reasoning and arguments. It seems that support theory can incorporate many of the facets expounded by the other models mentioned previously.

What most models mentioned so far share is their "internalist" stance that explains overconfidence and the hard-easy effect in terms of different biases or distortions and less than

perfect processing of information. They also present a rather bleak picture of the human being as being rather poor in adjusting to changing realities. Several training programs designed to overcome these biases reported in the literature have only reported modest improvements and the improvements they do report do not seem to generalize from one knowledge domain to another (Allwood & Granhag, 1999; Keren, 1991; Lichtenstein et al., 1982).

Externalist models

The ecological models draw on the ideas by May described above and suggest explanations in terms of biased, non-representational selection of items typically used in general knowledge tasks. The error models, on the other hand, suggest explanations in terms of regression effects due to random error components. These two perspectives have been brought together in several hybrid ecological/error models (e.g. Juslin et al., 2003; Soll, 1996; Suantak et al., 1996). The internalist models (Thurstonian) and the ecological models (Brunswikian) have been referred to as being on either side of a distinction between two modes of uncertainty which have been inspired by "…two of the great probabilists in the history of psychology, namely L. L. Thurstone and Egon Brunswik" (Juslin & Olsson, 1997, p. 345).

Ecological models

Thurstonian uncertainty is described as being located inside us humans much in the same sense that has been delineated by the internalist models presented earlier (e.g. the stage model, the process model). In contrast to this, Brunswikian uncertainty is external in the sense that judgement errors and faulty decisions arise due to less than perfect correlations between subjectively known data and uncertain aspects of the world. This means that no improvement in the reliability of our information processing system can alleviate this source of error.

Inspired by Brunswik, and independently of each other, Gigerenzer, Hoffrage and Kleinbölting, 1991 and Juslin, 1994 (reported in McClelland & Bolger, 1994) produced remarkably similar ecological models. The basic assumptions of these models are that people are seen as (1) well adapted to their environments, (2) they accurately and with little conscious effort store information regarding the frequency of occurrence of events and (3) these stored frequencies are the basis for probability or confidence judgements. Thus Brunswiks' notion of people as "intuitive statisticians" means that people are frequentists.

If we use the same example as mentioned on page 12 in connection to the process model and suppose that a person tries to figure out which country has the larger population then, according to the ecological models, the reasoning might be along the following lines. A person

uses all European countries as the proper reference class and the populations of the same countries as the target variable. To help the person make a choice between the presented alternatives a probability cue must be used. This is defined as a variable related to the target variable, as in this example "larger countries tend to have larger populations". If this fails another cue might be generated. The ecological models claim that if one compares the sizes and populations of all European countries in pairs, the ecological validity of this particular cue would be, perhaps 0.7. That means that it would work on 70 % of the occasions. They further assume that these ecological validities are cognitively represented in the individual as cue validities which they report conform to the reported confidence judgements, if, and this is important, the task is representative of a defined reference class. They argue that overconfidence is simply an effect of an informal selection of general knowledge tasks that are non-representative of the reference class in the environment.

Another prediction of these models is that if two sets of items, hard and easy, are generated from the same sampling process then the hard-easy effect should disappear. An investigation into this matter done in 1996 concluded that the ecological model's explanation of the hard-easy effect "…is falsified by the experimental results and analysis we have presented" (Suantak et al., 1996, p. 219). Suantak et al. (1996) also stated that "…the concept of objective cue validities as a basis for subjective probability judgements is fundamentally ambiguous" (p. 220). Note that informally selected, but previously used, items were used in the experiment in this paper. However, Cosmides and Tooby (1996) have shown that when using the "medical diagnosis problem" (see Appendix A) the often reported cognitive biases such as "base-rate neglect", "overconfidence" and "conjunction fallacy" tend to disappear if the problem is expressed in a frequentist representation.

Error models

The proponents of the importance of the role of statistical errors in the data of calibration research have argued that overconfidence and the hard-easy effect may more or less disappear if one controls for scale-end effects and regression effects in the data (Erev, Wallsten, & Budescu, 1994; Juslin et al., 2000; Soll, 1996). Erev et al. (1994) exemplifies the role of error using three constructs: true judgement, an error distribution and a response rule. A true judgement, T, is the likelihood of choosing the correct answer in a 2AFC task *i. Ti* is estimated from 0 to 1 and assumed to be influenced by random error *e*. The degree of subjective confidence is translated into a numerical confidence judgement *r*. Thus we have $r_i = f(Ti + e)$.

Erev et al. (1994) further proposed that, "…on the very reasonable assumption that judgements have an error component associated with them, the possibility exists that the phenomena of over- and underconfidence are often or primarily statistical consequences of how the data have been analyzed" (p. 523). However, they are not saying that overconfidence is necessarily or entirely statistical artefacts. The position they are arguing for is that the relation between subjective probability and actual accuracy in a particular context "…needs to be established after controlling for random factors in judgement or response" (Erev et al., p. 523). Juslin et al. (2000) came to a similar conclusion.

Recent developments

The models briefly described above still exert considerable influence on present day research in several ways. Many of the core ideas about how to explain miscalibration in individuals are built upon and further refined to explain a wider range of data. A major challenge facing the researchers who want to develop a "new" model is that they need to be able to explain precisely the phenomena which were inexplicable by the earlier models.

The *ecological rationality* research program presented next is an attempt to explain several of the cognitive biases identified in previous research, not solely by internal processes, but rather in connection with their ecological rationality, that is, their being useful adaptations to the environment.

Ecological rationality research program

In 2003 Todd and Gigerenzer presented a research program for studying simple decision heuristics inspired by H. A. Simon's ideas about human rationality. In 1981 Simon proposed (reported in Todd & Gigerenzer, 2003) that human rationality is constrained by two unrelated sets of bounds, external (e.g. the cost of search for information) and internal (memory constraints or information processing speed). In this research program these sets of bounds are seen as related and "may fit together like blades in a pair of scissors" (Todd & Gigerenzer, 2003, p. 143). From this perspective the internal bounds of the cognitive system can be shaped, for instance by evolution, to take advantage of the structure of the environment. In this sense, human beings exhibit *ecological rationality* and make good enough decisions by exploiting the external information structures in the environment. The authors argue that this is a rather positive view on the human being as a decision maker in contrast to other researchers who portray humans as suffering from cognitive illusions, irrationality and being cognitive misers unable to adapt to varying circumstances.

Todd and Gigerenzer (2003) claim that "…less information and processing can actually enable greater accuracy than more in some cases." (p. 145). This is probably counterintuitive to many but the authors report some studies where experts (they don't mention what type of experts) have been shown to base their judgements on "surprisingly few pieces of information" (Todd & Gigerenzer, 2003, p. 154). Hence their focus on "fast and frugal heuristics".

The authors propose that there are many different types of heuristics that are thought to make up parts of what they term the *adaptive toolbox*. This is seen as a collection of specialized cognitive mechanisms such as search for information, stopping-rules for search and decision-making that evolution and learning have built into the human mind. The building blocks of this adaptive toolbox may be put together to form a variety of fast and frugal heuristics and next we briefly outline one of them which, according to the authors, have received most attention.

Ignorance-based decision making. Some simple heuristics actually rely on a lack of knowledge to make appropriate decisions. For instance, the recognition heuristic can be used to recognize faces or names but lack of recognition can also be used in making a decision. If there are many items one doesn't recognize and therefore is ignorant of, Todd and Gigerenzer (2003) suggests that one disregards those items and simply chooses what one recognizes. These heuristics work in parallel.

This has been experimentally tested in relation to making an investment portfolio on the stock market. Laypersons where asked to form a portfolio based on ten companies they recognized and their portfolios were compared to professional fund managers. In this experiment done in 1996-1997 by Borges, Goldstein, Ortmann and Gigerenzer (cited in Todd & Gigerenzer, 2003) the ignorance-driven recognition heuristic outperformed highly trained fund managers who used all available information, as well as randomly formed portfolios, which sometimes beat the experts.

The authors argue that by studying ecological rationality one may "go beyond the widespread fiction that basing decision making on more information and computation will always lead to more accurate inferences" (Todd & Gigerenzer, 2003, p. 160). They give an example concerning children's language acquisition where the cognitive limitations actually seem beneficial.

The restrictions of the developing mind enables accurate learning of only a fraction of the environment, which then provides a scaffold to guide subsequent learning about the environment in an adaptive way. In sum, the above account argues that simple heuristics have a selective advantage over more complex cognitive strategies. This does not mean that humans don't use more complex cognitive processes, but under many circumstances it might be more beneficial or cost-effective to use some simpler rules of thumb one has learnt to rely on.

Multi-factorial, multi-level open systems model?

This is mainly a theoretical construct, but based on conclusions drawn from reviewing lots of empirical studies. It could be viewed as a "presumably false yet formally highly probable non-empirical statement" (Popper, 1963, p. 336). At this point I would like to make a digression which will serve as a major point that I would like to make in connection with making models and building theories.

I think it's virtually impossible to construct a grand general model that is able to explain a wide range of empirical phenomena over many domains of knowledge. It is much more likely that one succeeds in formulating a theory and a model which is more limited in scope, defined to accommodate a range of phenomena in a well specified domain and this could probably be done to a higher degree of accuracy. I base this on Popper's studies of the contents of theories which he contrasts with the calculus of probability. To illustrate how informative content in a theory stands in opposition to the probability of the same theory Popper (1963) has given the following example:

> Let *a* be the statement "It will rain on Friday", *b* the statement "It will be fine on Saturday" and *ab* the statement "It will rain on Friday and it will be fine on Saturday": it is then obvious that the informative content of this last statement, the conjunction *ab*, will exceed that of its component *a* and also that of its component *b*. And it will also be obvious that the probability of *ab* (or, what is the same, the probability that *ab* will be true) will be smaller than that of either of its components. Writing Ct (a), for "the content of statement *a*", and Ct (ab) for "the content of the conjunction *a* and *b*", we have
>
> (1) $$Ct(a) < Ct(ab) > Ct(b).$$
>
> This contrasts with the corresponding law of the calculus of probability,
>
> (2) $$p(a) > p(ab) < p(b)$$
>
> where the inequality signs of (1) are inverted. Together these two laws, (1) and (2), state that with increasing content, probability decreases and *vice versa*, or in other words, that content increases with increasing improbability. (…)
>
> This trivial fact has the following inescapable consequences: if growth of knowledge means that we operate with theories of increasing content, it must also mean that we operate with theories of decreasing probability (in the sense of the calculus of probability).

> Thus if our aim is the advancement or growth of knowledge, then a high probability (in the sense of the calculus of probability) cannot possibly be our aim as well: *these two aims are incompatible.* (p. 295).

Here it's interesting to note that, in line with Popper, Allwood and Granhag (1999) has *not* suggested a theory or model but pointed out a multitude of factors that reasonably could affect confidence judgements under real-life circumstances and factually affect people in several professions. They have further argued against the notion that few-factor models adequately can accommodate a wide range of phenomena found in calibration research. Allwood and Granhag have analyzed a substantial part of previous articles on the subject and, based on that, presented a list of factors (selected from Table 7.2 on page 134) which may influence realism in confidence judgements: type of confidence judgement (retrospective vs. predictions; item-specific vs. frequency), type of knowledge (semantic, episodic, procedural, implicit), number of alternatives (one vs. many), type of elicitation (spontaneous vs. instructed), number of persons (individual vs. group), degree of experience (low vs. high), cost for search of information (low vs. high), temporal relation between task and decision (short vs. long), temporal relation between decision and outcome (short vs. long), stability of environment (low vs. high), feedback (yes vs. no), type of feedback (clear vs. ambiguous) and delivery of feedback (immediate vs. delayed). If one compares this list of factors with the explanations given by the previously mentioned models how to explain the lack of realism often found in calibration research one can immediately see that these models have only used a very limited number of factors. The message by Allwood and Granhag is that "no few-factor theory will do on a general level" (p. 142).

In a study by Jonsson and Allwood (2003) they further qualified the above mentioned message that it is especially true if distal and global factors are considered. These factors include, for instance, knowledge domain, gender and cognitive style. Proximal factors, such as the type of cognitive processes leading to a confidence judgement, could be affected by distal factors such as knowledge domain and cognitive style. Rozenblit and Keil (2002) showed that people are much more overconfident in the knowledge domains about technical apparatuses and natural phenomena than in almanac questions such as those often used in 2AFC GKQs task.

As an example, consider the research done by Murphy and Winkler in 1971 (cited in Allwood & Granhag, 1999) on meteorologists. They have been found to be very well calibrated in their predictions. An explanation of this may be that they have lots of experience in making forecasts, they have amassed on enormous amount of data to make retrospective comparisons and their feedback is clear and quite immediate (is it rainy and windy here today?). Compare this

with a lawyer who is working with a case the outcome of which is determined by a court six months in the future (temporal relation). While working s/he can affect the outcome in various ways, self-involvement is (probably) high, the type of feedback is clear but delayed and the degree of experience could vary greatly from lawyer to lawyer. A further comparison with one of the usual tasks given in calibration research could further highlight this matter. Consider the factors involved in a 2AFC GKQ task. This is a current judgement of your own semantic knowledge with only two options which you are instructed to perform. Self involvement is (possibly) quite low and often no feedback is given so you can't learn to improve your performance.

Taking heed of Poppers remark one would hope for the development of many more models of human judgement under uncertainty which are adapted to different real-life circumstances, since it's highly improbable that someone succeeds in building a general model that can account for all the various ways and contexts in which one may make a confidence judgement.

Another philosophical remark could be made at this point. There are those who think that there will always be an explanatory gap in between what we know of where different processes in the brain take place and what it's like to experience those very same processes (Scheele, 2002). Or, if we turn the tables, even though we might have a reasonably good understanding of what processes affect human decision making we might never fully understand how the brain realizes those processes. This is yet another challenge facing the development of a cognitive neuroscience of metacognition.

Previous research on frequency judgements

In comparison with the abundant research on singular confidence judgements, the reported studies on global assessments or frequency judgements is rather scarce. Nevertheless, I'll try to set the stage as to why frequency judgements have drawn some interest. Generally people reason differently about an individual case than about a set of cases, especially when the set is presented in the form of statistics.

If one reads in the newspaper about the mistreatment someone has received, rather than only being presented with statistical information, then this arouses more empathy or anger because it's a concrete person whom it is easy to identify with. Even doctors have been shown to give more expensive treatment when faced with a single patient. When the same doctor is confronted by a set of patients he is much more likely to recommend treatments that are in line with a stricter rational cost-benefit analysis (Griffin & Buehler, 1999).

At the same time this example shows the difference between rule-based versus exemplar-based reasoning (Juslin & Olsson, 2004). When people consider a single case they apparently

focus on information that is relevant to that case and neglect the rules of probability that links that case to broader categories or frequencies, also called base-rate neglect. Those who have access to aggregate information stored statistically usually engage in reasoning which invokes rules that are relevant to statistically-based properties of set inclusion. Gigerenzer has made the bold claim that "the effect of frequency representations and judgements on 'cognitive illusions' is the strongest and most consistent 'debiasing method' known today" (cited in Griffin & Buehler, 1999, pp. 49-50).

Cosmides and Tooby (1996) has in a similar vein argued that the notoriously difficult problem used within the "heuristics and biases" program for eliciting base rate neglect can be solved by most people rather easily when the percentages in the original "medical diagnosis problem" are given in frequencies instead. For yet another poignant point of view in this context, see the last sentence of the next section.

The confidence-frequency effect

The most intriguing finding concerning research on frequency judgements has been referred to as the confidence-frequency effect. This effect describes the relationship between confidence judgements (which typically result in overconfidence) and frequency judgements. When confidence judgements are overconfident then frequency judgements tend to show good realism, that is, they are fairly accurate. When confidence judgements exhibit good realism, then frequency judgements tend to result in more or less underestimation or underconfidence.
This confidence-frequency effect has been reported 1991 by Gigerenzer, Hoffrage and Kleinbölting (cited in McClelland & Bolger, 1994) and by Sniezek and Buckley (1991) among others.

Previous studies concerning the confidence-frequency effect are not conclusive. Whereas some researchers have been able to replicate this effect in their studies (e.g. Treadwell & Nelson, 1996) others have found that both measures show substantial overconfidence (e.g. Brenner et al., 1996). Griffin and Buehler (1999) after three studies succinctly concluded that, "The studies presented here imply that under most real-life circumstances, intuitive judgements are equally biased regardless of the level of aggregation or whether frequency and probability is used" (p. 75).

Explanations of the confidence-frequency effect

According to Gigerenzer et al., 1991, (see McClelland & Bolger, 1994) confidence and frequency judgements belong to different reference classes. Consider the 2AFC GKQ example given above on page 12, the reference class is European countries but the frequency judgement itself belongs

to the participant's own reference class of previously answered questions in a similar situation. In sets of informally selected items the frequency judgements should show perfect realism because people have the experience that general knowledge tests are more difficult than they seem and therefore give less optimistic frequency judgements. Another inference from this theory is that when representative randomly sampled item sets are used, then frequency judgements should exhibit underestimation because the item-specific confidence judgements on the same questions that made up this set would exhibit, more or less, perfect realism.

Next, let's briefly look at how Sniezek and Buckley's (1991) *dual-process hypothesis* explains this effect. From their studies they concluded that confidence judgements for single items are affected by retrieval and evaluation of information about the item content. Miscalibration can hence be attributed to faulty information processing. When frequency judgements are considered they suggest that participants don't evaluate the information in the same way because of (1) the large number of items and (2) the heterogeneous content of those items.

When making a global or frequency judgement participants presumably take themselves into consideration. They might estimate their expertise relative to the demands of the task or their previous performances on similar tasks, or of the time and effort allocation. By making this distinction between knowledge of the subject matter and knowledge about one's previous test performances, it is possible to understand the item-specific confidence judgement and the frequency judgement as different types of judgement, perhaps affected by different processes or at least different content and therefore they need not exhibit identical levels of confidence.

When Treadwell and Nelson (1996) performed two studies to examine the confidence-frequency effect they found that the dual-process account best described their data. They also added the suggestion that frequency judgements may exhibit better realism than confidence judgements.

A similar explanation as the dual-process hypothesis has been given by Griffin and Tversky in 1992 (reported in Griffin & Buehler, 1999), but they also suggested that it could also be due to the participants only considering those items of which they are certain when assessing their performance over a whole set of questions and fail to adjust that estimate upwards because they could in fact be correct on some of the items for which they have guessed. This suggestion was due to that Griffin and Tversky observed that frequency judgements in 2AFC problems sometimes are below 50 %. This goes against the notion that simply by chance one would expect an accuracy of 50 % (on average, over the long run).

Liberman (2004) took this suggestion seriously and made three experiments to explore if this failure to correct for guessing can account for the discrepancy between local and global

assessments of confidence, or in our terminology, confidence and frequency judgements. His results suggest that people don't make a correction for guessing unless specifically asked and this normative failure can explain the discrepancy between confidence and frequency judgements in calibration studies. He also took issue with the assumption that frequency judgements would be more realistic in general as compared with confidence judgements. Liberman states clearly that "…the lack of overconfidence in global estimates is not evidence of greater realism or normativeness but rather a product of, and in fact dependent on, a familiar normative failure – the failure to make allowance for random guessing" (Liberman, 2004, p. 731).

Absolute versus relative frequency judgements

Treadwell and Nelson (1996) reported that frequency judgements not always exhibit good realism or underestimation. Likewise, Brenner et al. (1996) and Liberman (2004) found that relative frequency judgements (i.e., estimate given in per cent) can show overestimation. Hoffrage et al. (2002) have proposed that there might be a difference in how people make an absolute versus a relative frequency judgement. They made the claim that it is more cumbersome to make an estimate in relative frequencies as compared with natural frequencies or absolute frequencies (i.e., estimate given in natural numbers). To my knowledge, no one has made a study to explore if there in fact is a difference in realism between absolute and relative frequency judgements when 2AFC GKQs are used.

Summary and a remark

The models presented above have more or less successfully attempted to explain the general overconfidence phenomenon typically reported for confidence judgements and the degree of realism found for frequency judgements in calibration research. Another source of contention has been how to explain the hard-easy effect. Taken together, the enduring question has been whether these phenomena reflect genuine psychological phenomena or if they depend on methodological consequences.

A more recent finding is what is referred to as the confidence-frequency effect, which in short denotes that the level of confidence judgements typically is higher than the corresponding frequency judgement. This effect is something of a paradox since people can maintain a high confidence across single item knowledge claims but still estimate that their general performance level over the same set of items is much lower. Based on this curious finding it has been proposed that when people make these metacognitive judgements they do so depending on at

least partly different processes or content (e.g. Allwood & Granhag, 1999; Sniezek & Buckley, 1991).

The distinction between confidence and frequency judgements is related to a schism between subjectivists or Bayesians and frequentists as reported by Johansson (2004). Bayesians maintain that probability should be seen as a subjective measure of belief and allow the assignment of probabilities to unique events and require these assignments to obey the probability axiom, that is, when considering the probabilities for different alternatives these probabilities should be additive and sum up to 1 (Kahneman & Tversky, 1996). Frequentists, who tend to be the same researchers as those subscribing to ecological models, on the other hand, interpret probability as long-run relative frequency and refuse to assign probability to unique events. Apart from this disagreement they depart in their views on whether overconfidence reflects a genuine psychological bias due to imperfect information processes or reflects specific tasks used by the researcher.

After this general synopsis of some of the main questions within the calibration research tradition, the following section is intended to present different aspects of realism and how it is measured by the calibration methodology. The following presentation pinpoints a few of the most relevant issues from an excellent description made by Johansson (2004).

Calibration research and methodology

Participants who partake in a calibration research study are usually asked to choose which answer alternative in a 2AFC GKQ set they believe to be correct and then to confidence judge the correctness of their choice. This is usually done on a half-range (50-100 %) scale. The participants receive instructions that say that 100 % stands for being absolutely certain of having chosen the correct answer, while 50 % means that they have made a guess. In the case of frequency judgements, after having answered a block of GKQs and confidence judged all items one by one, the participants are asked to estimate how many of the questions in the very same block of questions they have answered correctly. Thus, the "…realism in confidence judgements may generally be defined as the extent to which confidence judgments conform to the proportion of correct assertions or *accuracy*" (Johansson, 2004, p. 5). A concrete example of this is that if a person states that s/he is 70 % certain in the chosen answer and have assigned it the corresponding probability 0.7 then, in the long run to be judged as realistic, it means that on all of the questions assigned the same probability value, 70 % of these questions should be correct.

This general definition of realism in confidence judgements also applies to frequency judgements. "The degree of realism in people's frequency judgements is defined as the extent to

which their assessments of overall accuracy conform to their actual accuracy." (Johansson, 2004, p. 5). There are other more specific measures such as calibration, over/underconfidence and resolution that are used by researchers to describe the relation between people's confidence judgements and accuracy. These measures of realism are called calibration measures and are typically used to analyze the realism in participants' confidence judgements. They are presented more fully in the next section.

Calibration measures

The following text in this section is cited from a study made by Allwood, Granhag and Johansson (2003) since the present author does not think that he can improve on that description.

> *Calibration* reflects the overall relation between the level of confidence judgements and the accuracy. A calibration score of 0 indicates perfect calibration and higher values reflect poorer realism. The formula for computing calibration is:
>
> (1) $$\text{Calibration} = 1/n \sum_{t=1}^{T} n_t \, (r_{tm} - c_t)^2$$
>
> In (1), n is the total number of questions answered, T is the number of confidence classes used, c_t is the proportion correct for all items in the confidence class, r_t, n_t is the number of times the confidence class r_t was used and r_{tm} is the mean of the confidence ratings in confidence class r_t. Thus, calibration is computed by first dividing participants' confidence ratings into a number of confidence classes. Next, for each confidence class, the difference is taken between the mean confidence for the items and the proportion of correct items. Finally, the squared differences multiplied by the number of responses in the confidence class are summed over confidence classes and divided by the total number of items.
>
> The *over/underconfidence* measure (henceforth called overconfidence) indicates that a person is overconfident (positive value) or underconfident (negative value). Overconfidence is computed in the same way as calibration, except that the differences are not squared. Higher absolute over/underconfidence values indicates higher over- or underconfidence, or less realistic confidence judgements. A value of zero indicates perfect realism.

Resolution

Loosely speaking, resolution reflects the ability of the participants to distinguish, by their confidence ratings, between two sets of answers: one set that is correct and one set that is incorrect. The formula for computing resolution is:

$$\text{Resolution} = 1/n \sum_{t=1}^{T} n_t (c_t - c)^2 \quad (2)$$

Here, in (2), c is the proportion of all items for which the correct alternative was selected. A higher value reflects better resolution than a lower. (p. 550)

For more extensive descriptions of these measures and their derivation from the Brier score see the expositions by Keren (1991) and Lichtenstein et al. (1982). To achieve good resolution a person has to assign, for example, lower confidence to all questions answered incorrectly compared with the questions answered correctly. A concrete example might explain this further. Consider two persons A and B both achieving c = 75 %. Person A consistently states 75 % confidence on all answers whereas person B is 50 % and 100 % correct when being 50 % and 100 % confident. Consequently person B shows much better discrimination ability than person A. Even though only person B shows perfect resolution, both person A and B are equally realistic in terms of calibration and overconfidence (e.g., Goldman, 1986; Johansson, 2004). In addition to these measures of realism this study also use mean accuracy, mean confidence, mean frequency and frealism in analyzing the results of the experiment. Frealism measures the degree of realism in the frequency judgements. In this paper the over/underconfidence measure will be denoted as ouconfidence.

Hypotheses

The first hypothesis concerns the *confidence-frequency effect*. Based on the results from previous research the first hypothesis was that there would be overconfidence in the confidence judgements when the frequency judgements result in good realism and when the confidence judgements show good realism the frequency judgements should show underestimation.

The second hypothesis delves deeper into the issue of frequency judgements. Since previous researchers have suggested, but not investigated, that relative frequency judgements as compared with absolute frequency judgements may exhibit overestimation, that result was predicted.

The third hypothesis hinges on Liberman's study (2004) in which he proposed that the level of relative frequency judgements will approach the level of realism in the confidence judgements when participants receive an instruction to incorporate a correction for guessing, that is, to include in the correct answers half of those items on which they evaluated they have guessed the answer. Since the lack of correspondence between the confidence and the frequency judgements could have other explanations as well, the third hypothesis predicted only a partial replication of Liberman's results in the relative condition. This means that in the absolute format there was still expected a larger discrepancy between the confidence and the frequency judgements even though the participants received an instruction to incorporate a correction for guessing.

Finally, the fourth hypothesis predicted realistic frequency judgements after the difficult set of questions and underconfidence in frequency judgements after the easy set of questions in those conditions where the participants did not receive an instruction to make a correction for guessing. Hopefully there will be the same difference between the absolute and relative format as predicted by the second hypothesis. This hypothesis was based on the dual-process hypothesis explained earlier (p. 23) and it also follows from the hard-easy effect.

For the harder set of question one would presumably take into account that it felt harder, was answered a bit slower and so forth, and thus adjust one's estimate to a more realistic level and the confidence judgements should show overconfidence. For the easy set of questions one would also adjust downwards from the level of confidence judgements according to the confidence-frequency effect, which now should show better realism, but this would lead to an underestimation when making the frequency judgement.

Since none of the reviewed models or theories has made any comparison between the factor instruction for guessing and difficulty level of the questions I couldn't do anything but speculate that perhaps there would be a shift in the levels of the frequency judgements so that after receiving the instruction to make a correction for guessing then the level of the frequency judgements for the set of easy questions (where the level of the frequency judgements without correction for guessing is likely to show underconfidence) would be a bit more realistic whereas for the set of hard set of questions they should show overestimation.

Method

Participants

In all, 112 undergraduate students of psychology from Lund University, Sweden, participated in the study (71 women and 41 men). The participants' mean age was 27 years (range of 18 to 50

years). There were 28 participants who acted in each condition and they were randomized across 4 conditions. They were not given any reward or other gratifications for participating.

Design

This study was planned to investigate whether there is a difference in realism in frequency judgement when participants are asked to make (1) a frequency judgement with or without correction for guessing, (2) an absolute versus relative frequency judgement and (3) for easy and hard questions. The participants were randomized into four between-subject conditions. The conditions differed as to the type of frequency judgement the subjects were asked to perform.

In condition 1 the participants were asked to perform an absolute frequency judgement without correction for guessing. In condition 2 the participants were given the same instruction as in Liberman's (2004) experiment 3, the neutral condition (Instruction, see Appendix B), to make a correction for guessing before they were asked to perform an absolute frequency judgement. In condition 3 the participants were asked to perform a relative frequency judgement without correction for guessing. Finally, in condition 4 the participants were given the same instruction as in Liberman's study 3, the neutral condition, to make a correction for guessing before they were asked to perform a relative frequency judgement. In all conditions there were 80 general knowledge questions (GKQs) which were divided into an easier and a harder set of 40 questions each. The order of easy-hard and hard-easy questions was balanced within each condition.

Materials

Questionnaire. A total of 80 two-alternative forced choice (2AFC) general knowledge questions (GKQ's) were given to the participants. All of these questions have previously been used by Allwood, Granhag and Johansson (2003), but questions which in that study had proven to be very easy (90 % or more correct) or very difficult (25 % or less correct answers in previous research) were eliminated. The original list of 90 GKQs was thus reduced to 80. The questions cover topics including nature, society, history, geography and lexical knowledge.

Procedure

The participants were recruited on a voluntary basis from courses given at the Lund university's department for psychology and were tested in smaller groups of 2-12 individuals at each time. Before the actual experiment was commenced, a small pilot study was conducted with four participants, to check how they understand the instructions and to get an approximate estimate

about the test completion time. On the basis of this pilot test there was no reason to change any of the instructions since they appeared to be understood as intended. The time used by these four participants to complete the test was in between 20 to 30 minutes.

In each condition the participants were asked to make a confidence judgement immediately after each question on a half-range-scale from 50 % (guessing) to 100 % (absolutely certain that the chosen answer is correct). It was pointed out that one of the two alternatives always was correct.

After each set of 40 questions the participants were asked to make a frequency judgement according to the condition-specific instructions concerning their total performance on those 40 questions. Then in all conditions the exact same procedure was repeated with the next set of 40 questions according to the same specific instructions as given for the first set of questions. All participants spent 15 to 30 minutes to complete the test.

Results

First I report the basic descriptive statistics for the five dependent measures, accuracy, confidence, calibration, ouconfidence and resolution. Next I report *t*-tests testing if there were any concrete order effect of the order of the easy and the hard questions on any of these measures and of the extent to which the measures calibration, ouconfidence and resolution differed from zero. Thereafter I present the results of the analysis of the effects of the between-subject factors Instruction: No correction for guessing/Correction for guessing and Format: Absolute/Relative and the within-subject factor Difficulty Level: Easy/Hard for which a 3 way ANOVA was computed. Next, under the heading *frequency judgements*, I present the results that are directly concerned with the core of my hypotheses.

Accuracy, confidence and the calibration measures

Table 1 (below) shows the means and standard deviations for the sets of easy and hard questions for each condition and each of the five dependent measures, accuracy, confidence, calibration, over/underconfidence (ouconfidence) and resolution.

As a first background analysis, the effect of the concrete order in which the participants encountered the two sets of 40 2AFC GKQ's was analyzed by means of a paired-sample *t*-test for each of the five dependent measures. Thus, for each dependent measure the results for the order easy-hard questions were compared with the results for the order hard-easy questions. The results showed that there were no significant differences on any of the dependent measures.

Table 1. Means and Standard Deviations (SD) of the Dependent Measures Accuracy, Confidence, Calibration, Ouconfidence and Resolution for the Easy Set and the Hard Set of Questions for Each Condition (n = 28 for Each Column)

	Condition			
	No correction for guessing		Correction for guessing	
Measure	Absolute	Relative	Absolute	Relative
Accuracy Easy	.751 (.092)	.755 (.090)	.775 (.071)	.773 (.105)
Accuracy Hard	.474 (.089)	.501 (.115)	.530 (.100)	.531 (.137)
Confidence Easy	.687 (.075)	.714 (.065)	.716 (.068)	.720 (.083)
Confidence Hard	.648 (.069)	.664 (.065)	.673 (.077)	.676 (.075)
Calibration Easy	.027 (.018)	.036 (.020)	.028 (.017)	.032 (.016)
Calibration Hard	.075 (.053)	.078 (.045)	.074 (.060)	.068 (.049)
Ouconfidence Easy	-.064 (.074)	-.040 (.087)	-.059 (.079)	-.053 (.085)
Ouconfidence Hard	.174 (.113)	.164 (.115)	.142 (.099)	.145 (.109)
Resolution Easy	.030 (.016)	.032 (.022)	.029 (.014)	.030 (.026)
Resolution Hard	.033 (.016)	.030 (.021)	.035 (.018)	.029 (.015)

Note: All means are given in proportions. Ouconfidence = a positive number indicates overconfidence and a negative number indicates underconfidence.

Next, I carried out one-sample *t*-tests to analyze for each condition and question difficulty level if the three calibration measures, calibration, ouconfidence and resolution differed significantly from 0 (test-value = 0, i.e., perfect realism for calibration and ouconfidence). The results for the condition No Correction for guessing/Absolute were for Calibration Easy, t (27) = 7.86, $p < .001$, Calibration Hard, t (27) = 7.42, $p < .001$, Ouconfidence Easy, t (27) = -4.52, $p < .001$, Ouconfidence Hard, t (27) = 8.15, $p < .001$, Resolution Easy, t (27) = 10.09, $p < .001$ and Resolution Hard, t (27) = 11.03, $p < .001$.

The results for condition No Correction for guessing/Relative were for Calibration Easy, t (27) = 9.43, $p < .001$, Calibration Hard, t (27) = 9.20, $p < .001$, Ouconfidence Easy, t (27) = -2.45, $p < .05$, Ouconfidence Hard, t (27) = 7.56, $p < .001$, Resolution Easy, t (27) = 7.76, $p < .001$ and Resolution Hard, t (27) = 7.53, $p < .001$.

The results for condition Correction for guessing/Absolute were for Calibration Easy, t (27) = 8.86, $p < .001$, Calibration Hard, t (27) = 6.51, $p < .001$, Ouconfidence Easy, t (27) = -3.92, $p < .001$, Ouconfidence Hard, t (27) = 7.57, $p < .001$, Resolution Easy, t (27) = 10.77, $p < .001$ and Resolution Hard, t (27) = 10.49, $p < .001$.

Finally, the results for condition Correction for guessing/Relative were for Calibration Easy, t (27) = 10.81, p < .001, Calibration Hard, t (27) = 7.40, p < .001, Ouconfidence Easy, t (27) = -3.32, p = .003, Ouconfidence Hard, t (27) = 7.05, p < .001, Resolution Easy, t (27) = 6.21, p < .001 and Resolution Hard, t (27) = 10.12, p < .001.

The results thus showed a significant difference on all tests, that is, all these calibration measures differed significantly from 0. Accordingly, the calibration and ouconfidence measures did not show perfect realism.

Evaluation of the effects of the factors instruction, format and difficulty level

Next, in order to analyze the effect of the factors Instruction, Format and Difficulty Level I carried out a 2 (Instruction: No Correction for guessing/Correction for guessing) x 2 (Format: Absolute/Relative) x 2 (Difficulty Level: Easy/Hard) mixed ANOVA on each of the five dependent measures. Note that these results are not of immediate relevance to our hypotheses because these are mainly concerned with the outcomes of the various frequency judgements.

The results showed a significant main effect of the between-subject factor Instruction (No Correction for guessing/Correction for guessing) for *accuracy*, F (1, 108) = 4.212, p < .05, η^2 = .04 but no significant effect on the other dependent measures. For the between-subject factor Format (Absolute/Relative) no significant main effects were found. There were no significant interaction effects of these two factors.

The results showed a significant main effect of the within-subject factor (Easy/Hard) for *accuracy*, F (1, 108) = 542.00, p < .001, η^2 = .83, *confidence*, F (1, 108) = 147.41, p < .001, η^2 = .58, *calibration*, F (1, 108) = 65.34, p < .001, η^2 = .38 and *ouconfidence*, F (1, 108) = 411.73, p < .001, η^2 = .79. However, there was no significant effect for *resolution*, which indicates that the participants' ability to discriminate between correct and incorrect responses was not affected by the difficulty of the questions.

Since no interaction effects were observed involving the within-subject factor Difficulty Level these results suggest that the outcomes do not depend on which condition the participants were in, but instead the results followed from the difficulty of the questions. After these general results we next deal with the analysis of the material that concerns the bulk of the hypotheses outlined above (pp. 27-28).

Frequency judgements

The frequency judgements, given at the end of each set of questions in all conditions, were transformed into relative frequencies (proportions) to enable an easy comparison with the

accuracy results (see Table 2 on page 35). For the set of easy questions the transformed mean frequency judgement were .495 for the No Correction for guessing/Absolute condition (actual accuracy, .751), .559 for the No Correction for guessing/Relative condition (actual accuracy, .755), .582 for the Correction for guessing/Absolute condition (actual accuracy, .775), and .721 for the Correction for guessing/Relative condition (actual accuracy, .773). For ease of comparison Figure 1 is presented below, which show the median and the distribution in quartiles for the measures confidence, accuracy and frequency judgements for the set of easy questions.

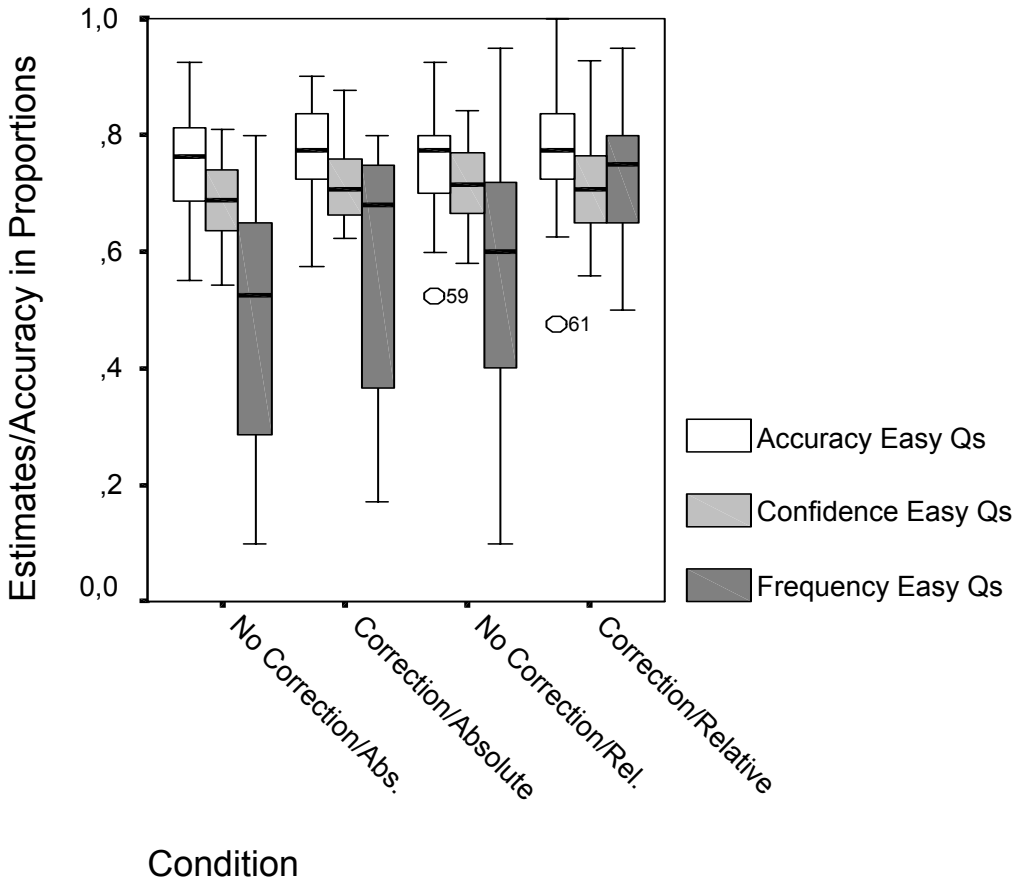

Figure 1. The median and the distribution in quartiles for the measures accuracy, confidence and frequency for the set of easy questions. The same convention as in Table 1 and Table 2 has been used therefore the scale on the y-axis is in proportions. Accuracy Easy Qs = accuracy for the set of easy questions. Confidence Easy Qs = confidence for the set of easy questions. Frequency Easy Qs = frequency judgements for the set of easy questions. The circles with the numbers next to them represent outliers. (n = 28 in each condition).

For the set of hard questions the mean frequency judgements were .427 for the No Correction for guessing/Absolute condition (actual accuracy, .474), .495 for the No Correction for guessing/Relative condition (actual accuracy, .501), .514 for the Correction for guessing/Absolute condition (actual accuracy, .530), and .652 for the Correction for guessing/Relative condition (actual accuracy, .531). For ease of comparison Figure 2 is presented

below, which show the median and the distribution in quartiles for the measures confidence, accuracy and frequency judgements for the set of hard questions.

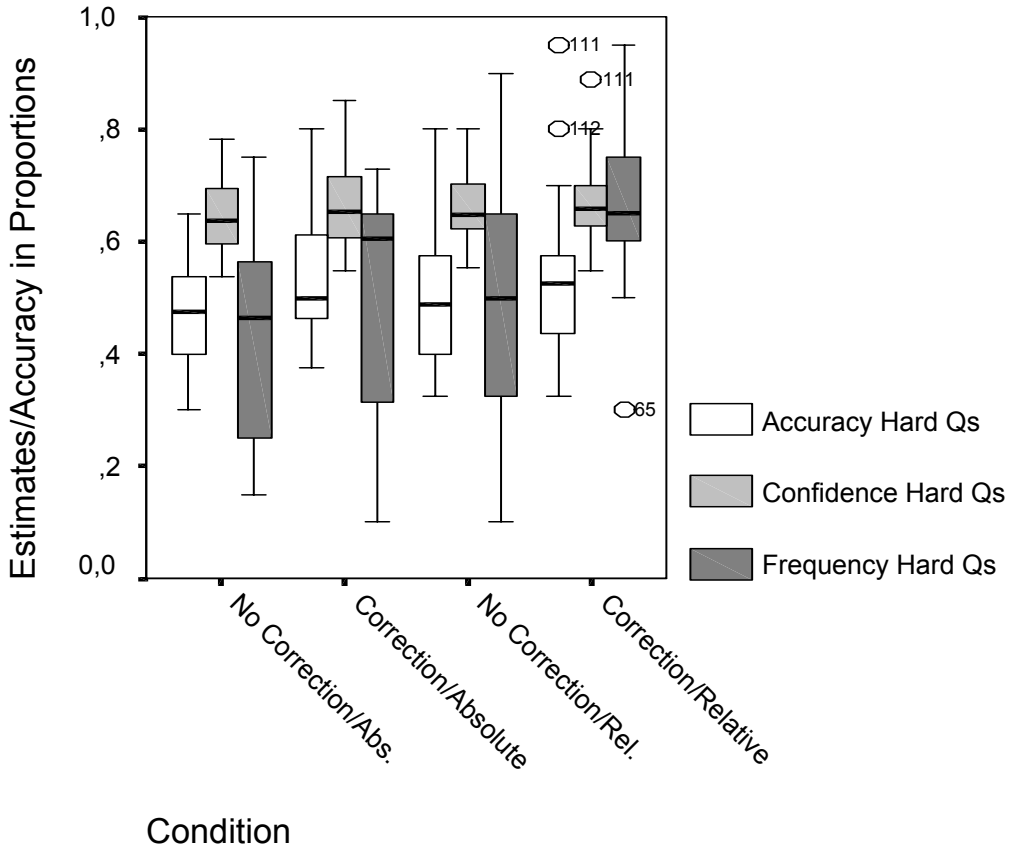

Figure 2. The median and the distribution in quartiles for the measures accuracy, confidence and frequency for the set of hard questions. The same convention as in Table 1 and Table 2 has been used therefore the scale on the y-axis is in proportions. Accuracy Hard Qs = accuracy for the set of hard questions. Confidence Hard Qs = confidence for the set of hard questions. Frequency Hard Qs = frequency judgements for the set of hard questions. The circles with the numbers next to them represent outliers. (n = 28 in each condition).

Next I computed one-sample *t*-tests to analyze for each condition and question difficulty level if the frequency judgements made by the participants differed significantly from 50 (test-value 50 = chance level for frequency judgements). The results were significant in all conditions for Frequency Easy, t (27) = all values < -1237.36, p < .001 and for Frequency Hard, t (27) = all values < -1234.02, p < .001.

The mean difference between the frequency judgements and the actual accuracy is shown in the lower part of Table 2 (below) for the different conditions and is called FRealism (standing for realism of the frequency judgements). The measure of realism which is presented in the table is based on the formula FRealism = Frequency Judgement – Accuracy, which was used to compute the results for both of the two sets of questions.

Table 2. Means and Standard Deviations (SD) of Frequency Judgements and Realism in Frequency Judgements, i.e. Frequency - Accuracy for Each Condition (n = 28/Condition/Column)

	Condition			
	No correction for guessing		Correction for guessing	
Measure	Absolute	Relative	Absolute	Relative
Frequency Easy	.495 (.212)	.559 (.211)	.582 (.197)	.721 (.110)
Frequency Hard	.427 (.177)	.495 (.212)	.514 (.182)	.652 (.124)
FRealism Easy	-.256 (.217)**	-.195 (.184)**	-.193 (.207)**	-.053 (.085)*
FRealism Hard	-.047 (.215)	-.006 (.199)	-.016 (.211)	.121 (.123)**

Note: For the FRealism measures * = $p < .05$ and ** = $p < .001$ refers to the comparison if FRealism in each condition differed from zero.

The results for the set of easy questions showed that 93 of the participants had a negative difference which means that their actual accuracy on answering the questions were higher than their estimated accuracy as reported by their frequency judgement. 17 of the participants had a positive difference which means that their frequency judgements overestimated their number of correct answers. Two of the participants showed perfect realism of their frequency judgements.

I carried out one-sample *t*-tests to analyze for each condition if the measure FRealism on the set of easy questions differed significantly from 0 (test-value = 0, i.e. perfect realism). The result for condition No Correction for guessing/Absolute was FRealism Easy, t (27) = -6.24, $p < .001$, for condition No Correction for guessing/Relative FRealism Easy, t (27) = -5.63, $p < .001$, for condition Correction for guessing/Absolute FRealism Easy, t (27) = -4.95, $p < .001$ and for condition Correction for guessing/Relative FRealism Easy, t (27) = -3.28, $p = .003$. The result of the same one-sample *t*-tests for the set of hard questions was only significant for condition Correction for guessing/Relative for FRealism Hard, t (27) = 5.17, $p < .001$. Within this condition all participants had a positive difference on the FRealism measure which means that they believed they chose the correct answer more often than they factually did.

Evaluation of the effects of the factors instruction, format and difficulty level on frequency judgements

After this, in order to analyze the effects of factors Instruction, Format and Difficulty Level, I computed a 2 (Instruction: No Correction for guessing/Correction for guessing) x 2 (Format: Absolute/Relative) x 2 (Difficulty Level: Easy/Hard) mixed ANOVA on the results for the frequency judgements and for the measures of FRealism. The results showed a significant main

effect of the between-subject factor Instruction (No Correction for guessing/Correction for guessing) for the *frequency judgements,* F (1, 108) = 14.03, $p < .001$, $\eta^2 = .12$ and for *FRealism,* F (1, 108) = 7.60, $p < .05$, $\eta^2 = .07$. The results also showed significant main effects of the between-subject factor Format (Absolute/Relative) for the *frequency judgements,* F (1, 108) = 9.64, $p < .05$, $\eta^2 = .08$ and for *FRealism,* F (1, 108) = 8.27, $p < .05$, $\eta^2 = .07$. Moreover, the results showed significant main effects of the within-subject factor Difficulty level (Easy/Hard) for the *frequency judgements,* F (1, 108) = 43.30, $p < .001$, $\eta^2 = .29$ and for *FRealism,* F (1, 108) = 229.92, $p < .001$, $\eta^2 = .68$. There were no significant interaction effects for these tests. This means there were simple main effects of all three factors Instruction, Format and Difficulty Level.

Discussion

The present study investigated whether there is a difference in realism in frequency judgements when participants are asked to make (1) a frequency judgement with or without a correction for guessing, (2) an absolute versus relative frequency judgement and (3) for easy and hard sets of questions. These factors were named: Instruction (No correction for guessing/Correction for guessing), Format (Absolute/Relative) and Difficulty level (Easy/Hard). All three factors were found to have a significant main effect on the level of the frequency judgements and their realism (frequency judgement – accuracy). In addition, there were no interaction effects. The results suggest that the level and the realism of frequency judgements are both affected by different factors. Below, I summarize and discuss these factors pertaining to the hypotheses. Before that I will make a comment in regard to the hard-easy effect and the overconfidence phenomenon.

The hard-easy effect and overconfidence

Since the issue as to why the hard-easy effect typically is found in calibration research was of no immediate interest in this study it may just be noted that this effect was present. From Figures 1 and 2 presented on pages 33-34 one can see that there was a marked overconfidence in all conditions for the hard set of questions and slight underconfidence in all conditions for the easy question sets.

Actually, this accord quite well with Keren's (1991) suggestion that participants use an anchor on a probability estimate reflecting intermediate difficulty which, for 2AFC GKQs is, 75 %. This anchor worked quite well for most of the participants in all conditions for the set of easy questions. However, as Keren suggested the participants did not adjust this anchor sufficiently when they encountered the set of hard questions. They did adjust a bit, which can be seen in Table 1 on page 31, but far from sufficiently.

Since neither interviews nor written protocols were used I can only speculate as to why the participants in all conditions exhibited such marked overconfidence on the hard question set. Perhaps some participants used a faulty cue as proposed by the process model and the ecological models. Perhaps some participants didn't consider the strength and weight of the evidence for the two alternatives or maybe they didn't consider evidence for their residual hypothesis as Tversky and Koehler (1994) has suggested. This would mean that some participants used the recognition heuristic and simply ignored the other alternative as Todd and Gigerenzer (2003) have proposed. Whatever is the case, let's now turn our attention to the issues pertaining to the factors investigated as mentioned above.

The confidence-frequency effect

The pattern of results more or less with three exceptions supported the first hypothesis that when the confidence judgements exhibit overconfidence then the frequency judgements result in good realism and when the confidence judgements show good realism then the frequency judgements show underestimation. On a general level, a comparison of the means for accuracy, confidence and frequency judgements regardless of the questions difficulty level shows that the confidence judgements were slightly overconfident and that the frequency judgements were underconfident.

If one separates the results for these measures for the sets of easy and hard questions then the confidence judgements for the set of easy questions show slight underconfidence whereas the frequency judgements show markedly more underconfidence. For the set of hard questions the confidence judgements showed overconfidence whereas the frequency judgements exhibited almost perfect realism.

More specifically, for the set of easy questions the level of confidence judgements was fairly realistic but showed slight and significant underconfidence in all conditions. The frequency judgements showed, as predicted, significant underestimation, but in the Correction for guessing/relative condition the frequency judgements were on the same level as the corresponding mean confidence judgements.

For the set of hard questions the level of confidence judgements showed significant overconfidence in all conditions while the level of frequency judgement was realistic in all conditions except in the Correction for guessing/relative condition where it was almost on par with the overconfidence exhibited by the confidence judgements. The levels of confidence and frequency judgements thus almost followed the predicted pattern.

These results are only partly compatible with Gigerenzer et al.'s 1991 (described in McClelland & Bolger, 1994) explanation of the confidence-frequency effect. Considering that I used sets of informally selected items then frequency judgements should show perfect realism but this was only true in three conditions for the set of hard questions. Furthermore, the frequency judgements showed underestimation in all conditions for the set of easy questions. This is what Gigerenzer et al. predicted would be the finding when representative, randomly sampled item sets are used. In this study underestimation occurred in spite of the fact that informally selected items were used.

The dual-process hypothesis seems to best account for the data with one notable exception. In all conditions, except the Correction for guessing/relative condition, on both levels of difficulty of the questions the mean level of the confidence judgements was always markedly higher than the comparable frequency judgements. Sniezek and Buckley (1991) made the prediction that frequency judgements would be lower than the confidence judgements and they also presented empirical support that the frequency judgements were more realistic, at least showed less overconfidence than the mean level of the confidence judgements. However, the dual-process hypothesis is falsified in the Correction for guessing/relative condition since both types of judgement show almost exactly the same level of underestimation for the set of easy questions and almost exactly the same level of overestimation for the set of hard questions.

Liberman (2004) suggested that if participants incorporate a correction for guessing in their frequency judgements they would approach the same level of realism as in the confidence judgements. The results in this study are partly compatible with Liberman's proposal. In the No correction for guessing/relative condition for both difficulty levels of the questions, the level of the frequency judgements was much lower than the mean level of the confidence judgements. In the Correction for guessing/relative condition for both difficulty levels of the questions the frequency judgements was almost on the same level as the mean level of the confidence judgements for both difficulty levels of the questions. The same type of comparison in the absolute format did not reveal the same type of pattern. According to the third hypothesis that was not expected either.

Absolute versus relative frequency judgements

The prediction of the second hypothesis, that relative frequency judgements as compared to absolute frequency judgements would exhibit overestimation, was not supported by the results except in the Correction for guessing/relative condition as compared with the Correction for guessing/absolute condition and there only for the hard set of questions. The participants in the

No correction for guessing in both absolute and relative conditions almost approached perfect realism in their frequency judgements for the set of hard questions which contradicts the hypothesis. However, the effect sizes for the factor format were quite small. These results suggest that one can expect good realism when people make frequency estimates on a set of hard questions and in terms of natural numbers. The suggestion made by Cosmides and Tooby (1996) that people make better estimates when asked to do so in natural numbers is therefore only given partial support since the participants in the present study gravely underestimated their performance on the set of easy questions even though they gave their estimates in natural numbers.

Although the results did not show overestimation for all relative frequency judgements there is, however, a pattern which is discernable from the presented results. For the format absolute/relative the main effect showed the following tendency. In all comparisons between the levels of absolute versus relative frequency judgements, within the same level of the instruction factor, the level of relative frequency judgements was always higher than the corresponding absolute frequency judgements. I coin the term *absolute-relative effect* to describe this pattern.

The results support neither the suggestion made by Brenner et al. (1996), nor the speculation by Hoffrage et al. (2002) that making an estimate in natural numbers would be easier and therefore should exhibit better realism. On the other hand, these researchers have not, to my knowledge, considered what effects the factors difficulty level and instruction could have on the level of absolute and relative frequency judgements. Liberman (2004) did consider the factor instruction (discussed next) but he did not consider the difficulty level of the questions or whether there would be a difference between the levels of realism in frequency judgements if they are given in terms of absolute or relative numbers.

The guessing instruction factor

The third hypothesis predicted a partial replication of Liberman's (2004) results in the relative format. To reiterate, when participants receive an instruction to incorporate a correction for guessing their level of relative frequency judgements should approach the level of realism found for the confidence judgements. This predicted pattern was only partly replicated in the present study in the relative format. When participants didn't receive an instruction to incorporate a correction for guessing before making a relative frequency judgement they made frequency judgements on a much lower level than the mean level of the confidence judgements. The participants who received an instruction to make a correction for guessing approached the same

level as the confidence judgements. This pattern is discernible from Figure 1 and Figure 2 on pages 33-34.

A significant difference in this study compared to Liberman's study 2, is that he reported only 3 of 134 participants whose mean estimate were below 50 %, that is, chance level, in his unrestricted (no instruction to correct for guessing was given) condition where they were free to estimate their performance. In comparison, in the No correction for guessing/relative condition in this study, there were 10 of 28 participants on the set of easy questions whose estimate was below chance level and 12 of 28 participants made estimates below 50 % on the set of hard questions. Liberman reported that none of the participants in his study 2 and 3 gave estimates below 50 % after receiving the instruction to correct for guessing and in this study there was only one who did. Strangely enough, this instruction to make a correction for guessing with the normative reminder that a performance level of 50 % would be expected by chance alone could not completely eliminate estimates below 50 % in the absolute format. In the Correction for guessing/absolute condition there were 9 of 28 estimates below 50 % for the set of easy questions and 10 of 28 estimates below 50 % for the set of hard questions.

This pattern of results contradicts the notion given by Todd and Gigerenzer (2003) and Cosmides and Tooby (1996) that people are frequentists and make by far better estimates when asked to do so in natural numbers. A not very likely, but possible, explanation of this is that the participants in the Correction for guessing/absolute condition did not understand the instructions. This explanation is not very likely since their confidence judgements and accuracy on the set of easy questions were fairly realistic and well above chance level and their confidence judgements for the set of hard questions actually showed overconfidence while their accuracy was almost on chance level.

In short, the factor Instruction did have a significant main effect, even though the effect sizes were small for this factor, on the level of the frequency judgements. The results herein only lend partial support to Liberman's suggestion that the often found good realism in frequency judgements is due to the normative failure that participants don't include a correction for guessing when making a frequency estimate. This is so since in the absolute format the level of frequency judgements did not approach the level of realism as in the confidence judgements.

Effects of the difficulty level of the questions

The prediction of the fourth hypothesis that frequency judgements would be more realistic after the set of hard questions and show underconfidence after the set of easy questions was partially borne out in both formats when the participants were not asked to make a correction for

guessing. The mean frequency judgements were realistic for the set of hard questions with the exception that they showed overestimation in the Correction for guessing/relative condition. The results for the set of easy questions showed significant underestimation in all conditions. It seems that the results follow nicely from the dual process account when at the same time one takes the hard-easy effect into consideration. It is noteworthy that the effect sizes for this factor, the difficulty level of the questions, were in the range of medium to large.

General discussion

To reiterate, since there were no interaction effects and only simple main effects of all factors on the level of the frequency judgements and for their realism, the results suggest that the level of realism in the frequency judgements is affected by different processes when participants are asked to make these metacognitive judgements, but the effect sizes show that the effects of instruction and format were small and the effect sizes for difficulty level of the questions were of medium to large size.

In the large, the dual-support hypothesis seems to best accommodate the data in the present study with one notable exception, that the levels of the confidence and frequency judgements given in the Correction for guessing/relative condition were essentially the same. Thus, the results indicate that it could plausibly be that participants make reference to different parts of their memory systems and different processes are involved when making the two kinds of investigated metacognitive judgements.

Since the effect of Liberman's instruction was only partially successful one must consider the possibility that there is some other factor or multiple factors involved which prevents the participants from making realistic frequency judgements in different conditions. Since I didn't interview the participants nor ask them to make written protocols wherein they could have described how and on what pieces of information they made their estimates, I can only speculate about which type of heuristics they used or failed to use.

The results show that the participants did not necessarily give more realistic frequency judgements when asked to do so in natural numbers since the estimates for the set of easy questions show more underestimation than the corresponding relative frequency judgements. This also suggests that there are other explanations as to why people are miscalibrated than just assuming that people are "natural frequentists" and have more difficulty making estimates in terms of relative frequencies. One explanation could be that the recognition heuristic worked quite well for most of the participants in all conditions for the set of easy questions. These items could perhaps be part of most participants' background knowledge as seen by their higher mean

level of accuracy. For the set of hard questions the items are obviously not to the same degree part of the participants general background knowledge (since the mean level of accuracy were much lower than for the set of easy questions) and therefore the participants might not have been able to recollect additional evidence for their chosen answer. If that was the case then they must rely more on the recognition heuristic or the familiarity of one answer alternative in order to choose an answer. This ignorance driven heuristic could have been involved for the set of hard questions since the results for actual accuracy were approximately on chance level in all conditions.

However, the design of this study was to investigate *if* the three factors studied had an effect on the level and the realism of the frequency judgements participants made. It was not designed to find out *how* or *why* these factors affected the frequency judgements. Furthermore the design did not aim to find out what type of reasoning procedures or processes the participants engage in when making confidence or frequency judgements. I propose that these matters should be seen as open empirical questions for subsequent research to investigate. Even though the results herein lend support to the conclusion that the three factors investigated has an effect, I suggest further replications (or partial replications) of this study. Since I have only been able to locate one previous study of the effect of instruction and none concerning the effect of using both an absolute or relative format when answering 2AFC GKQs, I especially recommend further exploration into these factors before one takes the effects of these factors as an established fact. The effects of the difficulty level of the questions have in comparison been quite thoroughly investigated.

Finally, to answer the question posed in the title which has guided most of the work presented in this paper, correction for guessing does improve realism in relative frequency judgements but only with respect to the set of easy questions. Correction for guessing for the set of hard questions had the effect that frequency judgements showed overestimation and no correction for guessing before relative frequency judgements showed almost perfect realism.

Correction for guessing using the absolute format improved realism slightly for the set of easy questions but for both levels of instruction the level of realism in frequency judgements showed significant underestimation. Correction for guessing using the absolute format did not make a significant difference on the set of hard questions in the absolute format since for both levels of instruction the frequency judgements only showed slight underestimation (they did not differ significantly from zero). In short, correction for guessing does only markedly improve the realism in relative frequency judgements on a set of easy questions.

Concluding remarks

Most of the accounts briefly sketched in the beginning of this paper involve explanations of metacognitive judgements in terms of different cognitive and methodological factors. Taken together, none of them could fully explain the results in the present study. Considering that the overwhelming majority of the reported studies do not use any task that one could expect that most people perform in real life situations one would hope that this would come more into focus in future research.

The fact that people don't always know what they claim to know has today become trivial. What is more interesting is how people come to trust someone as being knowledgeable. More specifically, on how many occasions does a person need to show that her/his estimates are fairly realistic before one concludes that s/he is trustworthy in a particular knowledge domain? Or should the question be posed differently? How many times can an expert fail to give realistic metacognitive judgements before s/he is degraded from the role of being an expert? A further complication of this matter is that people generally listen more to someone who seems to be confident in what they are saying than someone who does not express the same level of confidence.

Many previously reported studies have tried to explain why the hard-easy effect, the overconfidence phenomenon and the confidence-frequency effect occurs but not that many studies have tried to pinpoint the conditions under which people's metacognitive judgements in every day life could be expected to show good realism and therefore could be trusted regardless of the level of confidence expressed in these judgements. This would be a very welcome addition to our present state of knowledge about these matters.

Finally, since this paper will receive a very limited distribution and probably only will be read by a few psychologists and soon to be psychologists I would like to mention that when Oskamp did research on clinical psychologists in 1965 (reported in Allwood & Granhag, 1999) it was found that they were overconfident in diagnosing and predicting future behaviour. My question is, can clinical psychologists be assumed to have become more realistic in their judgements since then, and if not, what, if anything, can be done to increase their level of realism?

References

Allwood, C. M. (1995). Problematizing and its context. *Göteborg Psychological Reports, 25.* Göteborg: Göteborg University.

Allwood, C. M., & Granhag, P. A. (1996a). Realism in confidence judgements as a function of working in dyads or alone. *Organizational Behavior and Human Decision Processes, 66,* 277-289.

Allwood, C. M., & Granhag, P. A. (1996b). The effects of arguments on realism in confidence judgements. *Acta Psychologica, 91,* 99-119.

Allwood, C. M., & Granhag, P. A. (1999). Feelings of confidence and the realism of confidence judgements in everyday life. In P. Juslin & H. Montgomery (Eds.), *Judgement and decision making: Lens-modeling and process tracing approaches* (pp. 123-146). Hillsdale, NJ: Lawrence Erlbaum Press.

Allwood, C. M., Granhag, P. A., & Johansson, M. (2003). Increased realism in eyewitness confidence judgements: The effect of dyadic collaboration. *Applied Cognitive Psychology, 17,* 541-561.

Banich, M. T. (1997). *Neuropsychology. The neural bases of mental function.* Boston, MA: Houghton Mifflin Company.

Brenner, L. A. (2003). A random support model of the calibration of subjective probabilities. *Organizational Behavior and Human Decision Processes, 90,* 87-110.

Brenner, L. A., & Koehler, D. J. (1999). Subjective probability of disjunctive hypotheses: local-weight models for decomposition of evidential support. *Cognitive Psychology, 38,* 16-47.

Brenner, L. A., Koehler, D. J., Liberman, V., & Tversky, A. (1996). Overconfidence in probability and frequency judgements: A critical examination. *Organizational Behavior and Human Decision Processes, 65, 212-219.*

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgement under uncertainty. *Cognition, 58,* 1-73.

Dodson, C. S., & Schacter, D. L. (2002). When false recognition meets metacognition: the distinctiveness heuristic. *Journal of Memory and Language, 46,* 782-803.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12,* 83-87.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgement processes. *Psychological Review, 101,* 519-527.

Fernandez-Duque, D., Baird, J. A., & Posner, M. I. (2000). Executive attention and metacognitive regulation. *Consciousness and Cognition, 9,* 288-307.

Flavell, J. H., Miller, P. H., & Miller, S. (1993). *Cognitive development* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Freeman, W. J. (1999). *How brains make up their minds*. London: Orion Books.

Galotti, K. M. (1999). *Cognitive psychology in and out of the laboratory*. Pacific Grove, CA: Wadsworth.

Goldman, A. I. (1986). *Epistemology and cognition*. Cambridge, MA: Harvard University Press.

Griffin, D., & Buehler, R. (1999). Frequency, probability and prediction: Easy solutions to cognitive illusions? *Cognitive Psychology, 38,* 48-78.

Hoffrage, U., Gigerenzer, G., Krauss, S., & Martignon, L. (2002). Representation facilitates reasoning: what natural frequencies are and what they are not. *Cognition, 84,* 343-352.

Johansson, M. (2004). *Realism in metacognitive judgements: Effects of social factors*. Unpublished doctoral dissertation, Lund University, Lund.

Jonsson, A-C., & Allwood, C. M. (2003). Stability and variability in the realism of confidence judgements over time, content domain, and gender. *Personality and Individual Differences, 34,* 559- 574.

Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgement: A sampling model of confidence in sensory discrimination. *Psychological Review, 104,* 344-366.

Juslin, P., & Olsson, H. (2004). Note on the rationality of rule-based versus exemplar-based processing in human judgement. *Scandinavian Journal of Psychology, 45,* 37-47.

Juslin, P., Winman, A., & Olsson, H. (2000). Naïve empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. *Psychological Review, 107,* 384-396.

Juslin, P., Winman, A., & Olsson, H. (2003). Calibration, additivity and source independence of probability judgements in general knowledge and sensory discrimination tasks. *Organizational Behavior and Human Decision Processes, 92,* 34-51.

Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review, 103,* 582-591.

Kellogg, R. T. (1995). *Cognitive psychology*. London: SAGE Publications.

Keren, G. (1991). Calibration and probability judgements: Conceptual and methodological issues. *Acta Psychologica, 77,* 217-273.

Klayman, J., Soll. J. B., González-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes, 79,* 216-247.

Koriat, A. (2000). The feeling of knowing: Some metatheoretical implications for consciousness and control. *Consciousness and Cognition, 9,* 149-171.

Koriat, A., & Levy-Sadot, R. (2000). Conscious and unconscious metacognition: A rejoinder. *Consciousness and Cognition, 9,* 193-202.

Liberman, V. (2004). Local and global judgements of confidence. *Journal of Experimental Psychology: Learning, Memory and Cognition, 30,* 729-732.

Lichtenstein, S., Fischoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases* (pp. 306-334). Cambridge, UK: Cambridge University Press.

McClelland, A. G. R., & Bolger, F. (1994). The calibration of subjective probabilities: Theories and models 1980-94. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 453-482). New York: John Wiley & Sons.

Mussweiler, T., & Neumann, R. (2000). Sources of mental contamination: Comparing the effects of self-generated versus externally provided primes. *Journal of Experimental Social Psychology, 36,* 194-206.

Mussweiler, T., & Strack, F. (2000). Numeric judgements under uncertainty: The role of knowledge in anchoring. *Journal of Experimental Social Psychology, 36,* 495-518.

Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist, 51,* 102-116.

Olsson, H., & Juslin, P. (2000). The sensory sampling model: theoretical developments and empirical findings. *Food Quality and Preference, 11,* 27-34.

Popper, K. R. (1990). *The logic of scientific discovery* (14th ed.). London: Unwin Hyman Ltd.

Popper, K. R. (1963). *Conjectures and refutations.* London: Routledge & Kegan Paul.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: an illusion of explanatory depth. *Cognitive Science, 26,* 521-562.

Scheck, P., Meeter, M., & Nelson, T. O. (2004). Anchoring effects in the absolute accuracy of immediate versus delayed judgements of learning. *Journal of Memory and Language, 51,* 71-79.

Scheele, M. (2002). Never mind the gap: the explanatory gap as an artefact of naïve philosophical argument. *Philosophical Psychology, 15,* 333-342.

Schooler, J. W. (2002). Re-representing consciousness: dissociations between experience and meta-consciousness. *Trends in Cognitive Sciences, 8,* 339-344.

Shimamura, A. P. (2000). Toward a cognitive neuroscience of metacognition. *Consciousness and Cognition, 9,* 313-323.

Sniezek, J. A. & Buckley, T. (1991). Confidence depends on level of aggregation. *Journal of Behavioral Decision Making, 4,* 263-272.

Soll, J. B. (1996). Determinants of overconfidence and miscalibration: The roles of random error and ecological structure. *Organizational Behavior and Human Decision Processes, 65,* 117-137.

Suantak, L., Bolger, F., & Ferrell, W. R. (1996). The hard-easy effect in subjective probability calibration. *Organizational Behavior and Human Decision Processes, 67,* 201-221.

Todd, P. M., & Gigerenzer, G. (2003). Bounding rationality to the world. *Journal of Economic Psychology, 24,* 143-165.

Treadwell, J. R., & Nelson, T. O. (1996). Availability of information and the aggregation of confidence in prior decisions. *Organizational Behavior and Human Decision Processes, 68,* 13-27.

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

Tversky, A. & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101,* 547-567.

Yonelinas, A. P., Dobbins, I., Szymanski, M. D., Dhaliwal, H. S., & King, L. (1996). Signal-detection, threshold and dual-process models of recognition memory: ROCs and conscious recollection. *Consciousness and Cognition, 5,* 418-441.

Yonelinas, A. P. (2001). Consciousness, control and confidence: The 3 Cs of recognition memory. *Journal of Experimental Psychology: General, 130,* 361-379.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46,* 441-517.

Appendix A: The medical diagnosis problem

The original presentation of this problem was made by Casscells, Schoenberger and Graboys in 1978 (cited in Cosmides & Tooby, 1996) and replicated by Cosmides and Tooby (1996):

> If a test to detect a disease whose prevalence is 1/1000 has a false positive rate of 5 %, what is the chance that a person found to have a positive result actually has the disease, assuming that you know nothing about the person's symptoms or signs? ____% (p. 23)

Casscells et al. asked a group of faculty, staff and four-year students at Harvard Medical School to solve this problem (reported by Cosmides & Tooby, 1996). Only 18 % gave the correct answer, which is 2 %. A version of the same problem presented in a frequentist format is presented by Cosmides and Tooby (1996):

> 1 out of every 1000 Americans has disease X. A test has been developed to detect when a person has disease X. Every time the test is given to a person who has the disease, the test comes out positive (i.e., the "true positive" rate is 100 %). But sometimes the test also comes out positive when it is given to a person who is completely healthy. Specifically, out of every 1000 people who are perfectly healthy, 50 of them test positive for the disease (i.e., the "false positive" rate is 5 %). (. . .)
> How many people who test positive for the disease will *actually* have the disease? __ out of __ (p. 24)

When participants were given the information in this format, 76 % of them gave the correct answer.

Appendix B: Instruction

The following instruction from Varda Liberman's (2004) article was translated into Swedish:

> Logically your estimate should be at least 50 percent, because even people who make a random guess for each and every answer would, on average, get 50 percent correct. And, logically, if you knew the answer or had a good idea about the right answer on at least some of the questions your estimate should be higher than 50 percent because you should get more than 50 percent of those correct, plus (on average) half of the ones on which you just made a guess. However it is possible to be unlucky in one's guesses and to get less than 50 percent correct, and you are free to estimate less than 50 percent if you wish. What we want is your best estimate about your own particular performance on the 36 questions. (p. 731)

Note that the only thing I changed in the translation of this instruction was the number 36 to 40, since that was the number of questions in both of my sets of questions.